Wasserstein approximation of measures via orthogonal frames: Statistical and computational complexities

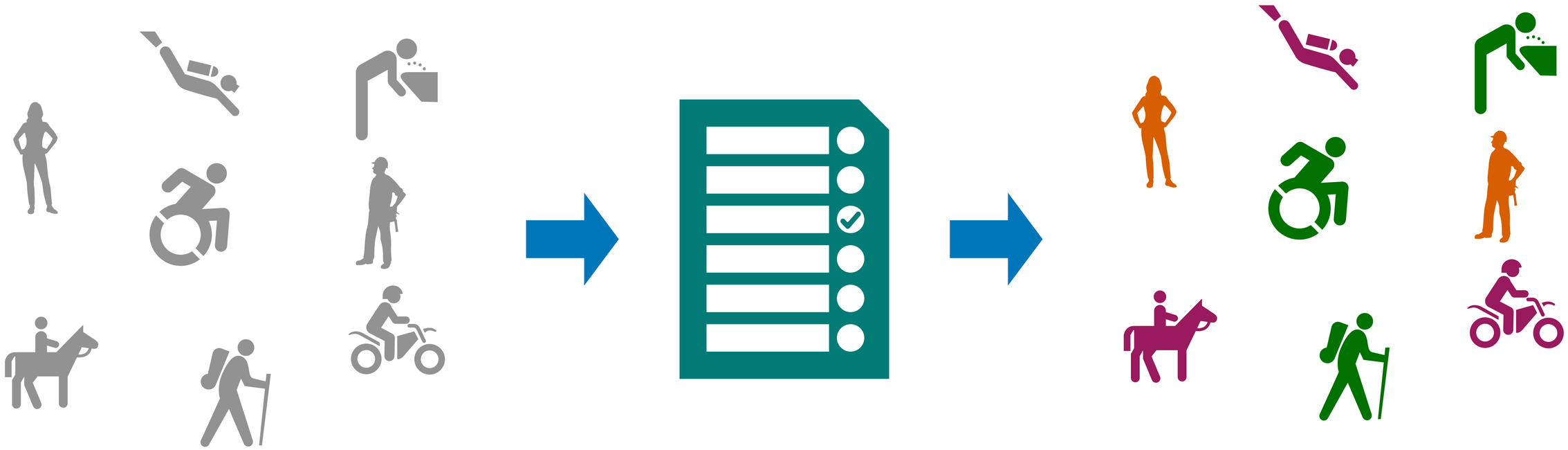
UN Encuentro de Matemáticas (70 años)
August 18th

Daniel López-Castaño
Johns Hopkins University

Joint with Mateo Diaz (U Chicago), Congwei Yang & Nicolás García-Trillos (U Wisconsin-Madison)

Motivation: A problem in Psychometry

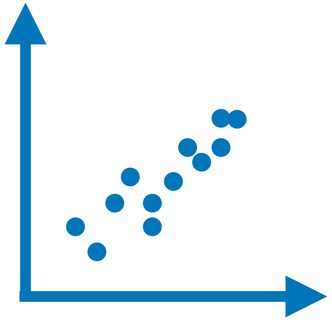
Psychometrics researchers aim to **understand personality types** through survey analysis.



The Mathematical Problem: Given $X \in \mathbb{R}^{d \times n}$, ***recover the labels*** $y \in \{1, \ldots, k\}$

Factor Analysis

Datasets with numerous variables can often be effectively represented using ***fewer latent variables (factors)***



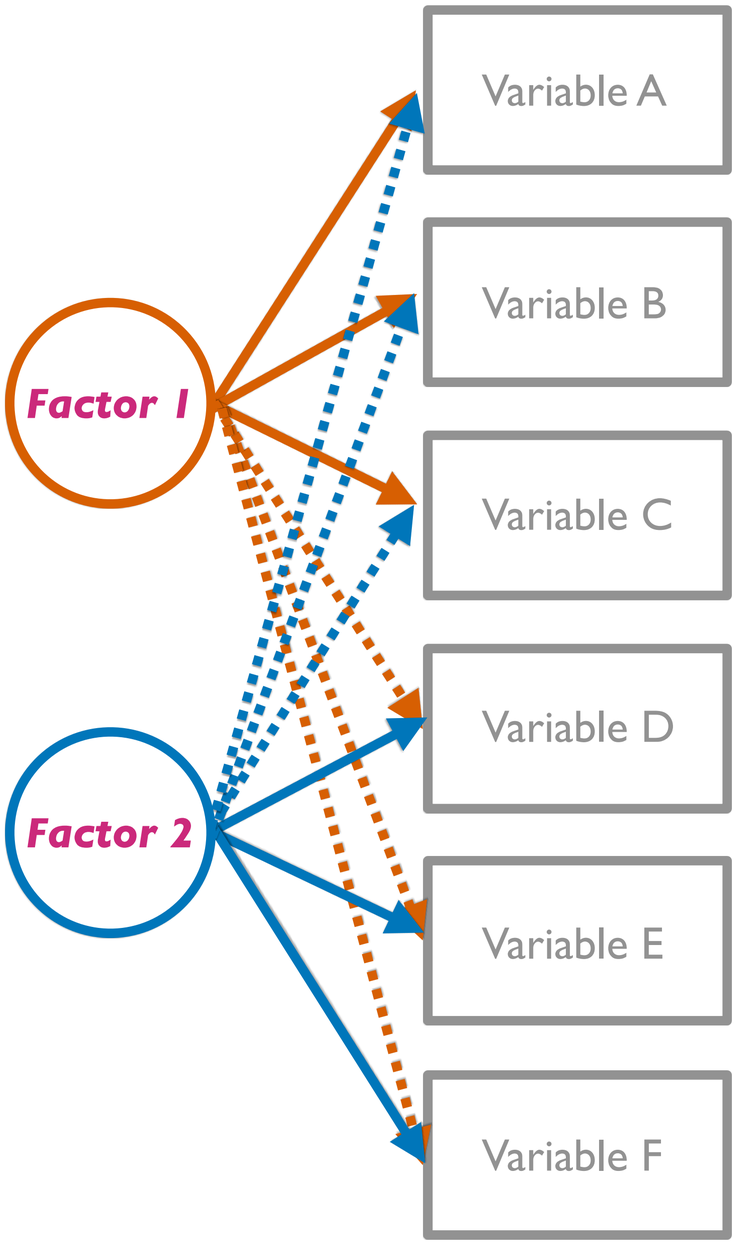
This is often described via a linear model:

$$X = FW^{\top}$$

Data

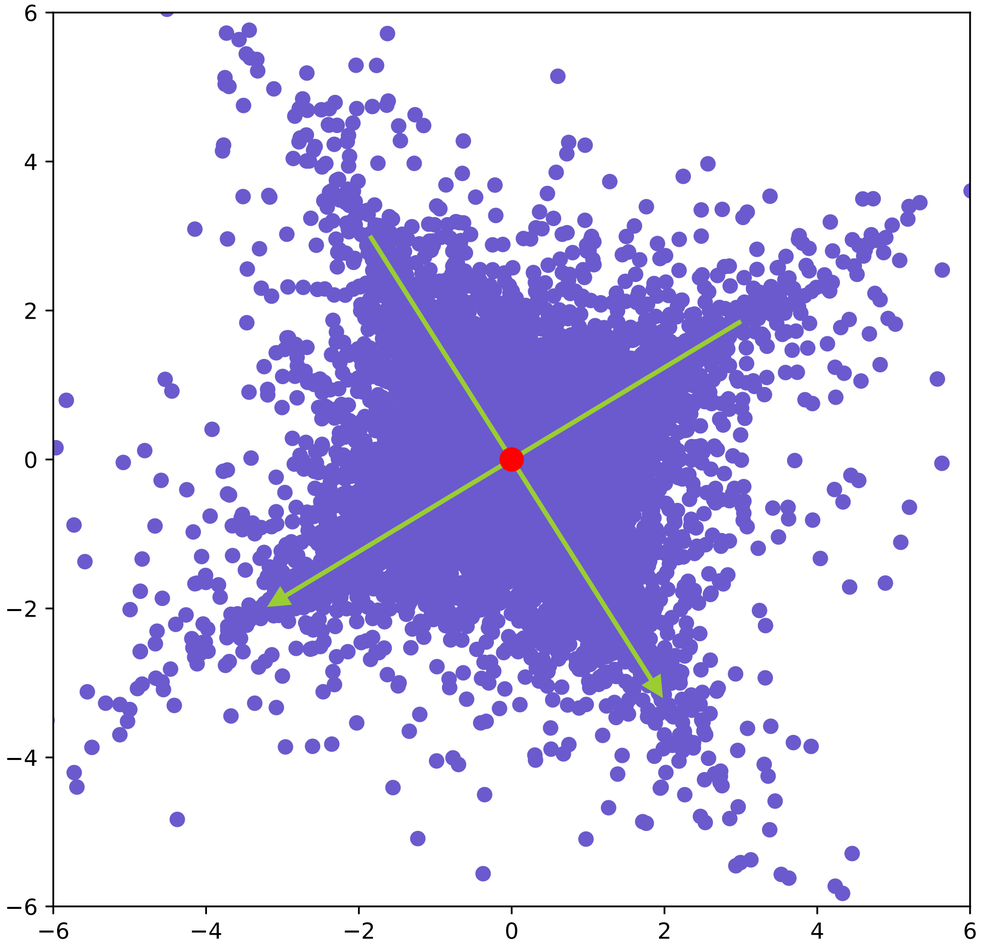
Loadings

Factors



Factor Analysis

Remarkably, in psychometrics, the data often exhibits ***radial structure***

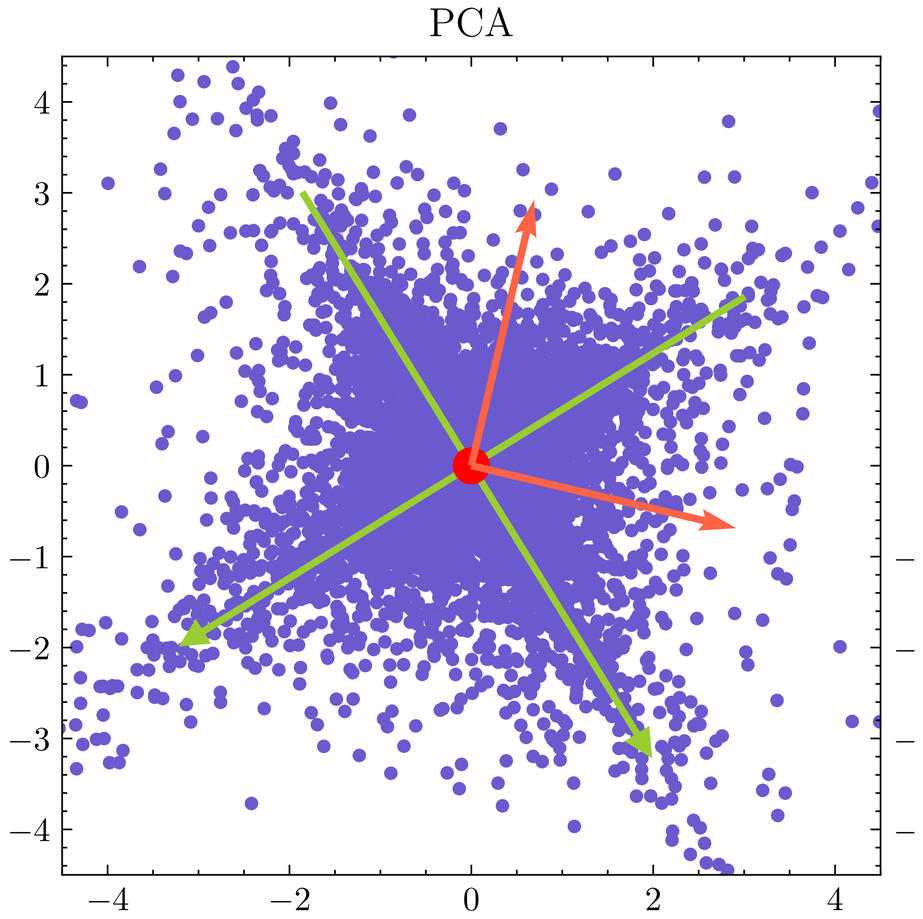


Intuitively, this means that the factorization looks like

$$X_i = Fw_i = \left[F_1, F_2, \ldots, F_k\right] \begin{pmatrix} 1 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_{k-1} \end{pmatrix}$$

and most information can be captured by **one factor!**

How to recover the radial structure?



Natural idea: Run PCA.

Problem: The covariance matrix is often close to a multiple of the identity.

$$\mathbb{E}\left[XX^{\top}\right] \approx \lambda\, Id$$

The **principal components** are essentially **random directions**!

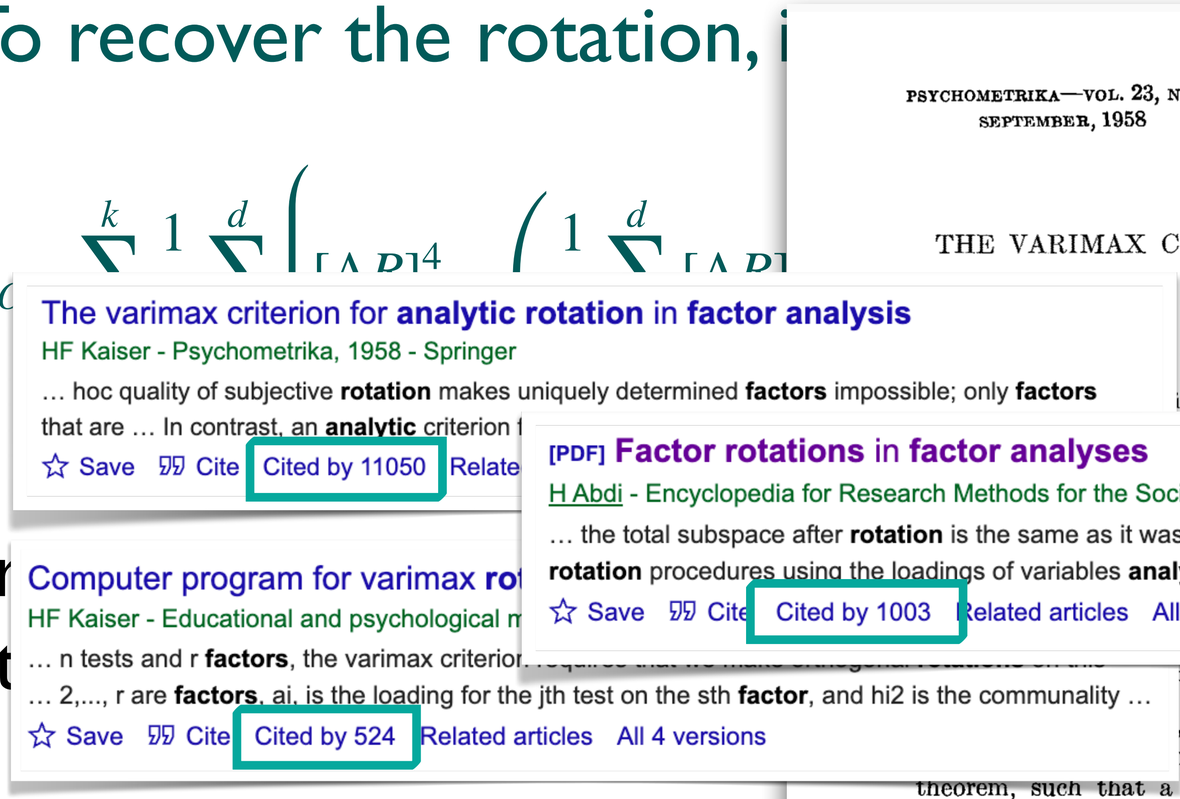
Population covariance is a diagonal matrix

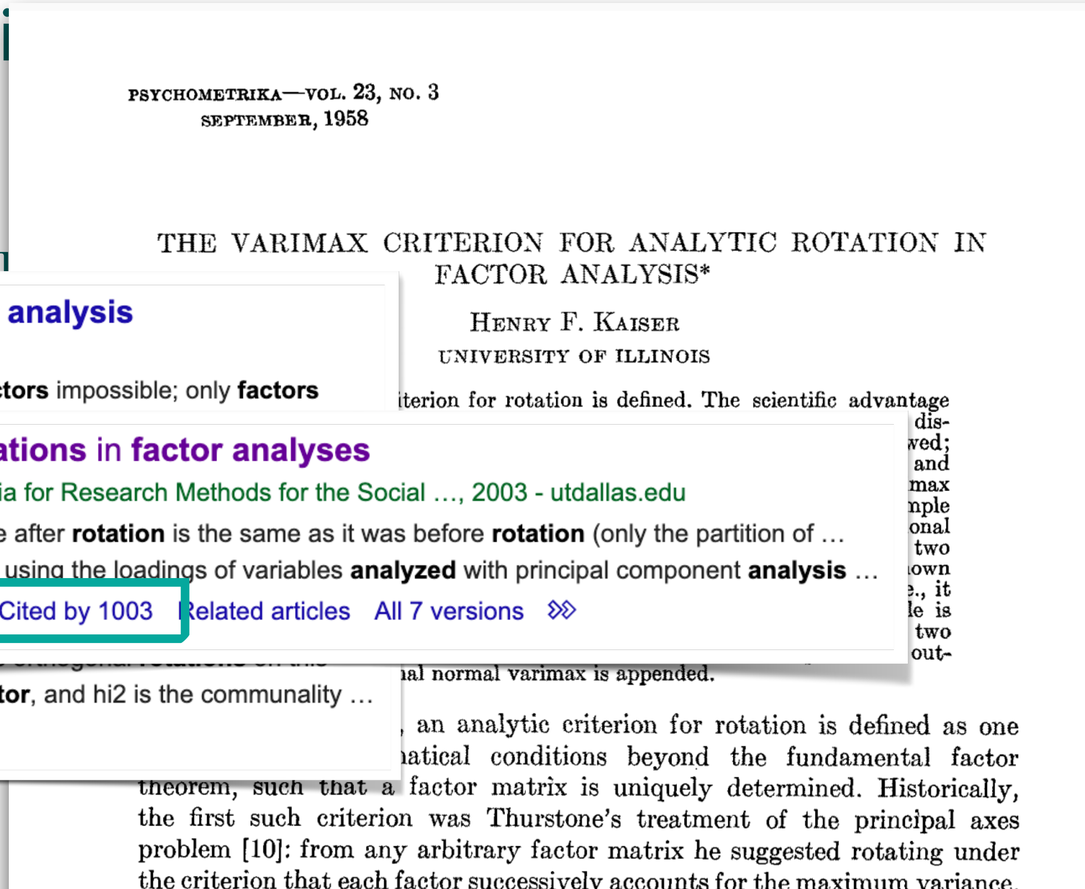
A classical solution: Varimax rotations

Clarify the relationship among ***factors***, by ***rotating*** the orthogonal basis found in PCA

To recover the rotation, i

$$R^\star \in \operatorname{argmax}_{R \in O}$$




PSYCHOMETRIKA—VOL. 23, NO. 3
SEPTEMBER, 1958

THE VARIMAX CRITERION FOR ANALYTIC ROTATION IN FACTOR ANALYSIS*

HENRY F. KAISER
UNIVERSITY OF ILLINOIS

…iterion for rotation is defined. The scientific advantage dis- …wed; and …max …mple …onal two …own …e., it …le is two out-

…al normal varimax is appended.

…, an analytic criterion for rotation is defined as one …matical conditions beyond the fundamental factor theorem, such that a factor matrix is uniquely determined. Historically, the first such criterion was Thurstone's treatment of the principal axes problem [10]: from any arbitrary factor matrix he suggested rotating under the criterion that each factor successively accounts for the maximum variance

Intuition: the r… determined by fit… moment.

How well does Varimax perform?

In theory (Rohe '22)

Under several assumptions on the distribution from which **factors** can be sampled:

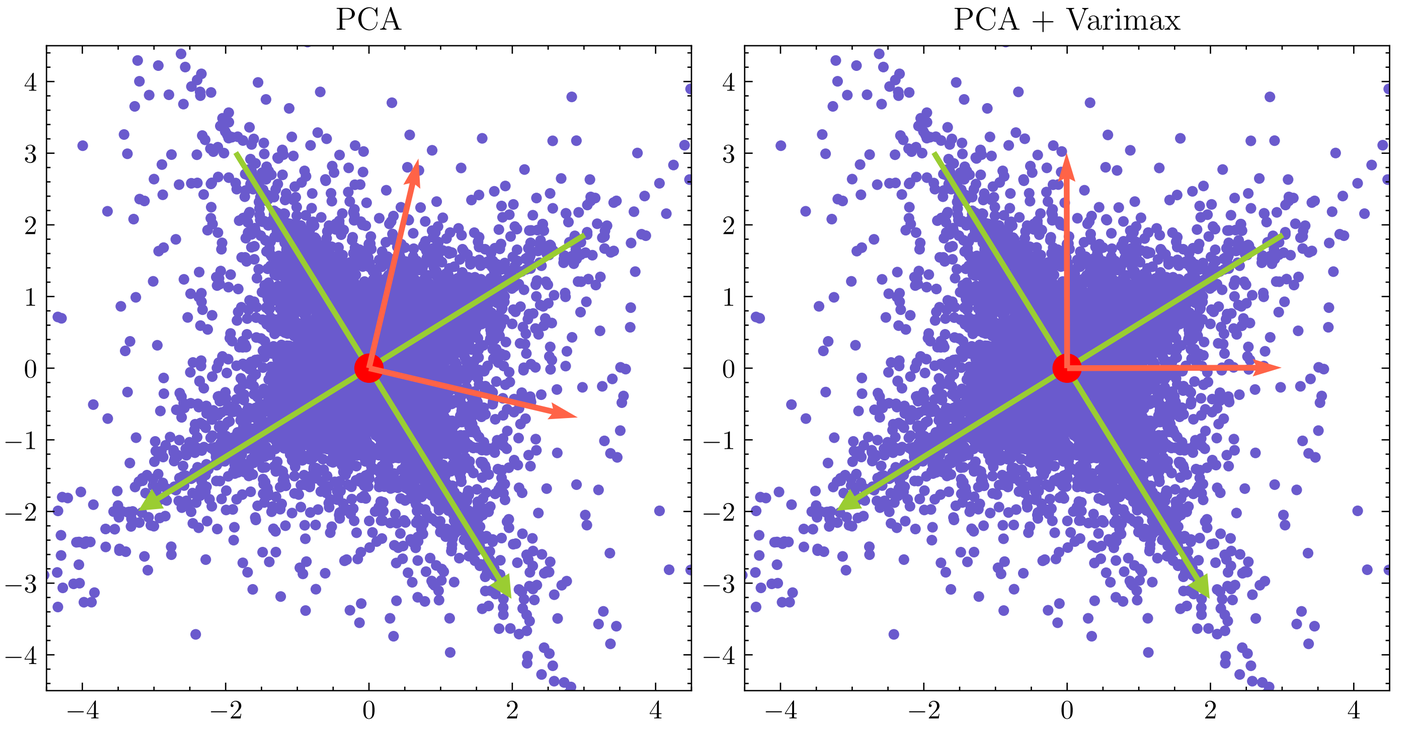
$n \times$ **data m**

$$\|\hat{X} - XP_n\|_{2\to\infty} \leq O_p(\Delta_n^{-0.24} \log^{3.75} n)$$

signed-permutation orbit of true factors

factor model

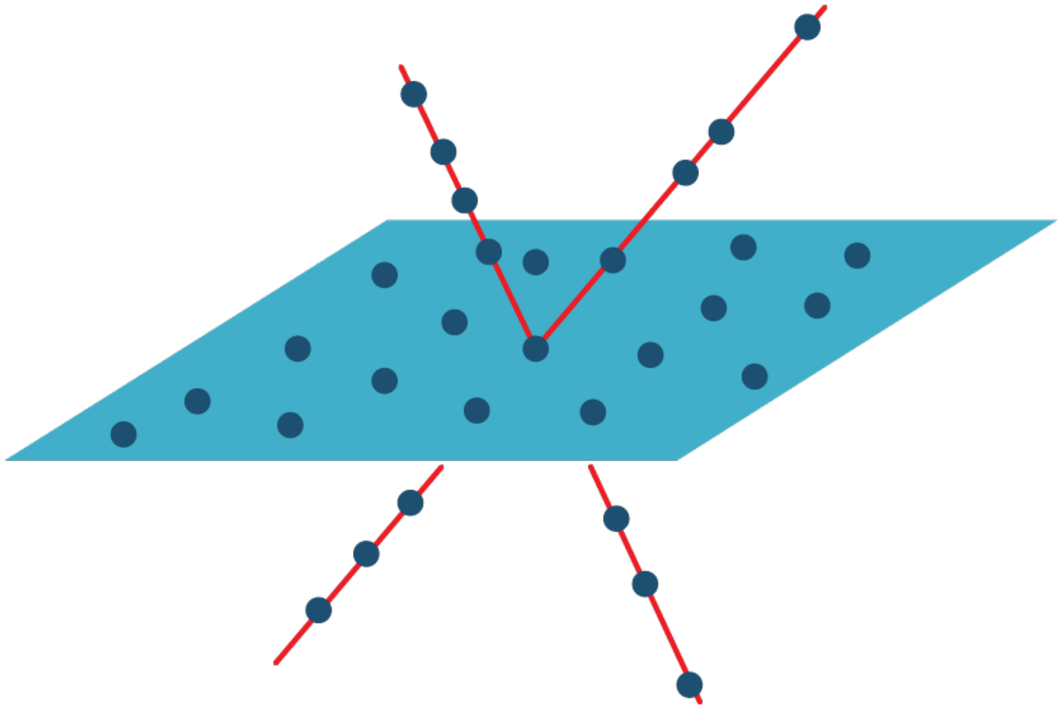
In practice:



Better alternatives?

Subspace clustering aims to recover a **union of subspaces** that capture **most of the data's information**

In particular, our problem is a special case of subspace clustering: consider k different 1-dimensional subspaces (**factors**)



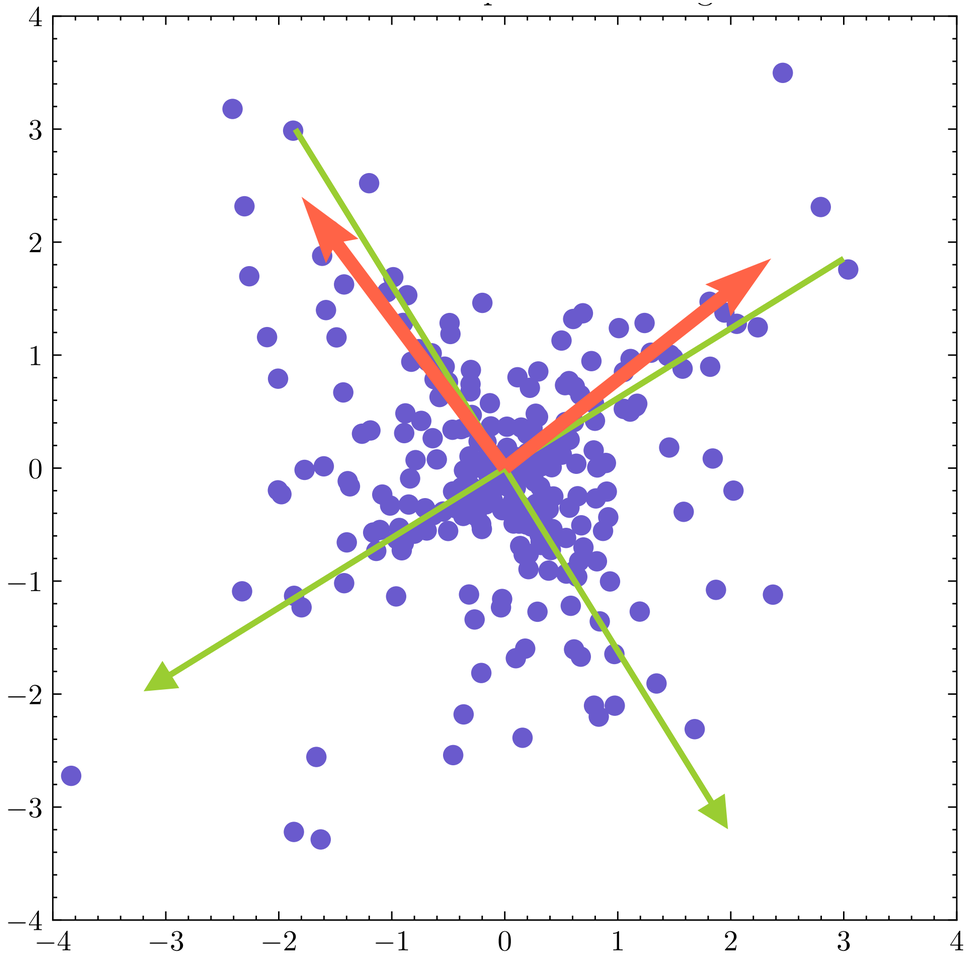
Note: Key applications in computer vision (Vidal, 2011)

Issues with subspace clustering

Issue 1: Most methods **lack robustness**

Issue 2: The most robust solution (Robust subspace clustering, Soltanolkotabi '14) requires **forming an n^2 similarity matrix** and **extracting top-k eigenvalues** on a second stage

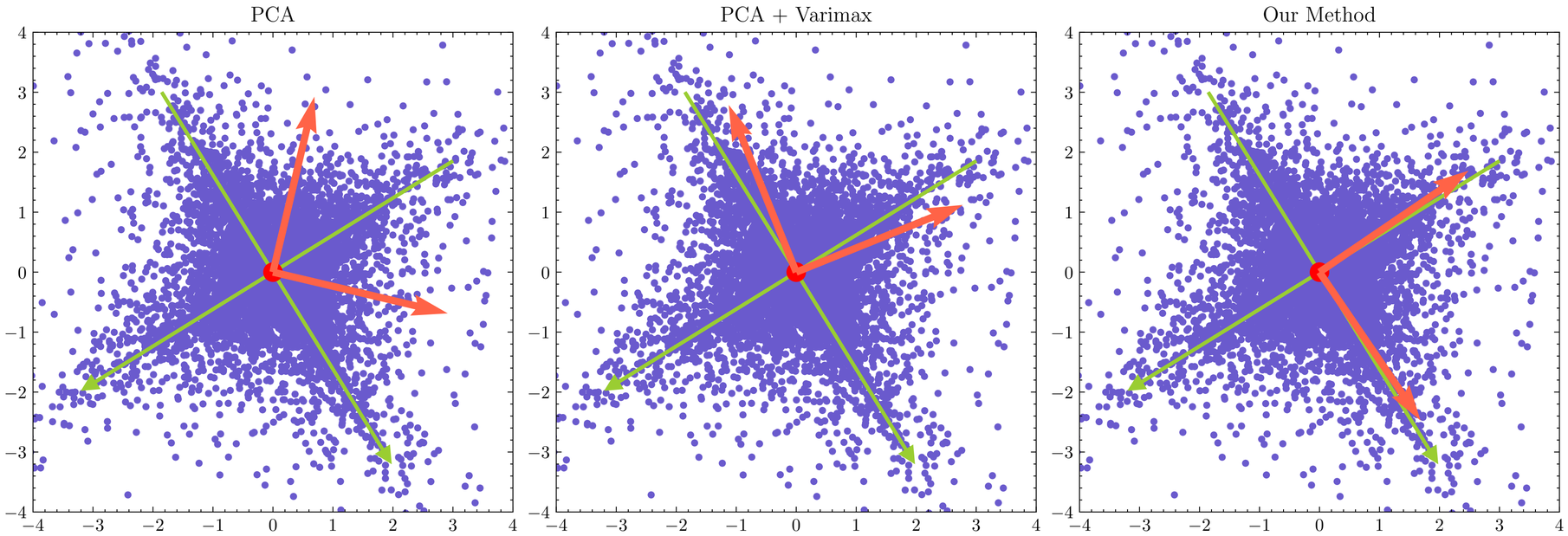
Issue 3: Subspaces are not constrained to be orthogonal to each other, thus it might allow **factors to correlate**



The question of the day

Can we have a scalable method that displays better statistical performance?

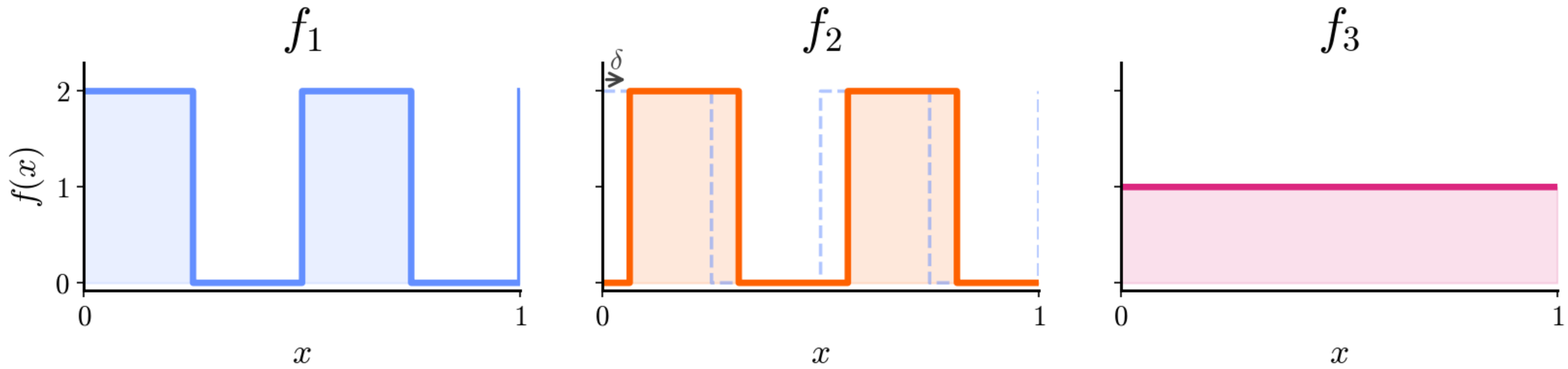
Yes! We introduce **Wasserstein Approx Measure (WAM) w/ manifold opt**



Agenda

- Motivation and previous work
- **Short Preliminaries**
- **Our method**
- **Statistical Guarantees**
- **Computational Guarantees**
- **Numerical Experiments**

Wasserstein Distance

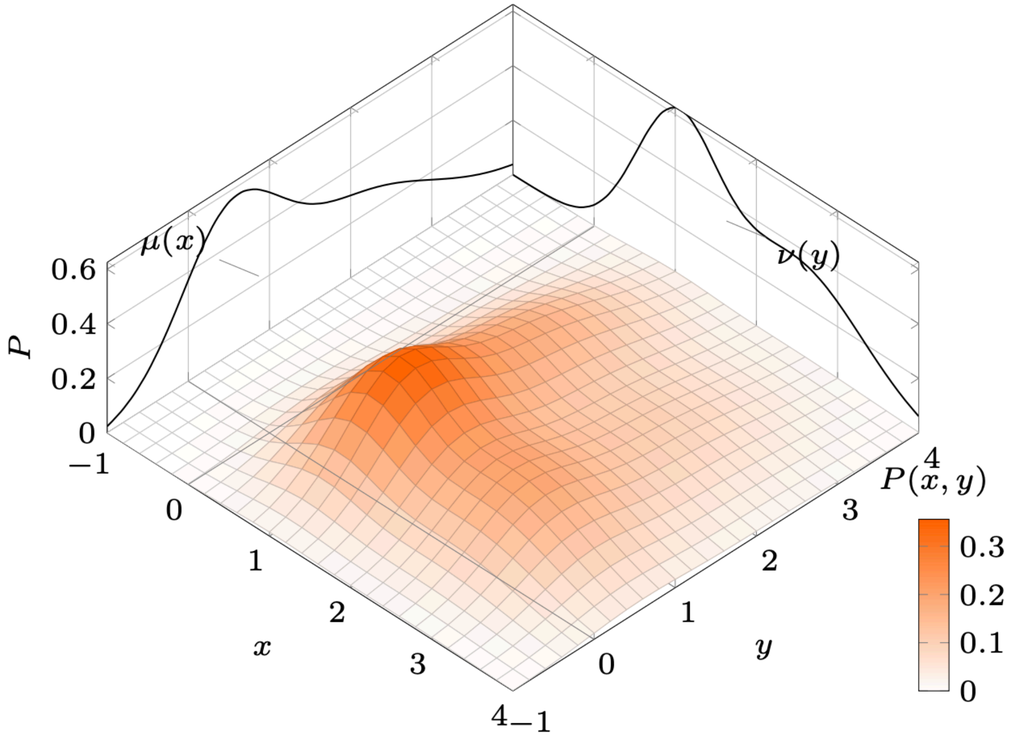


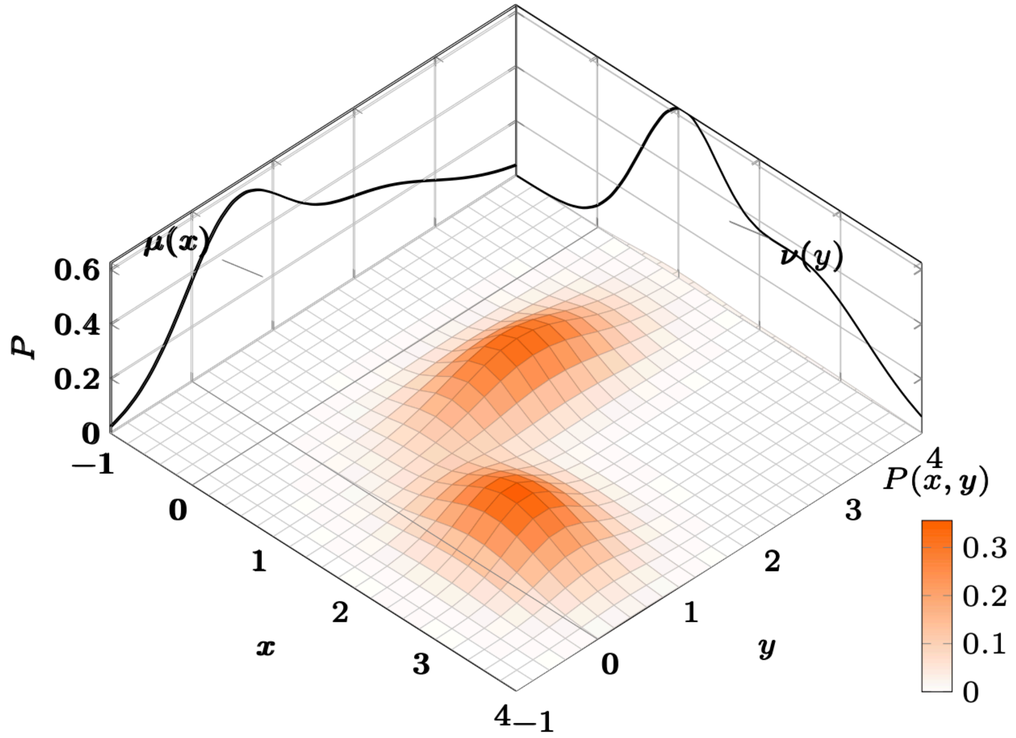
f_2 "should be" closest to f_1 than f_3, but in L_2-distance…

$$\|f_1 - f_2\|_2 = \|f_1 - f_3\|_2 = \|f_2 - f_3\|_2 = 1$$

Optimal transport makes it possible to define a **geometry** of a space of probability measures, and thus gives a definition of **distance** in this space.

$$\Pi(\mu,\nu) \stackrel{\text{def}}{=} \{P \in \mathcal{P}(\Omega \times \Omega) | \forall A, B \subset \Omega, P(A \times \Omega) = \mu(A), P(\Omega \times B) = \nu(B)\}$$





$$W_2(\mu,\nu) \stackrel{\text{def}}{=} \left(\inf_{P \in \Pi(\mu,\nu)} \iint d(x,y)^2 P(dx,dy) \right)^{1/2}$$

Wasserstein approx measures w/ orthogonal frames

Idea: The data is close to being **supported on a cross**

Let's find the closest distribution with that property!

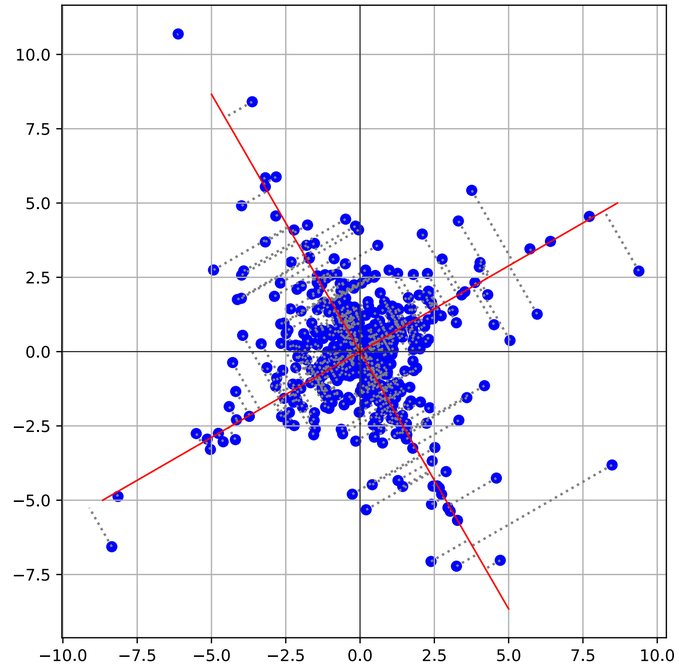
Considering

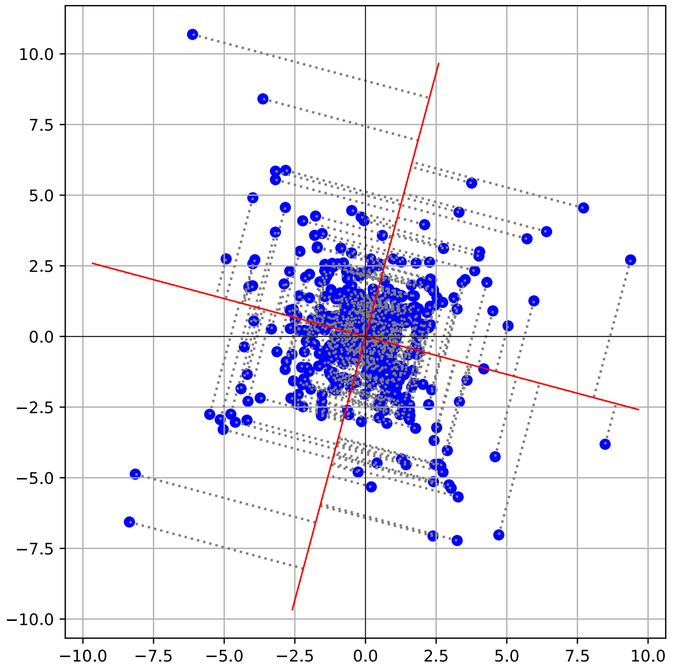
$\mathscr{P}_\times = \{\text{probability distributions supported on a cross with } k \text{ factors}\}$,

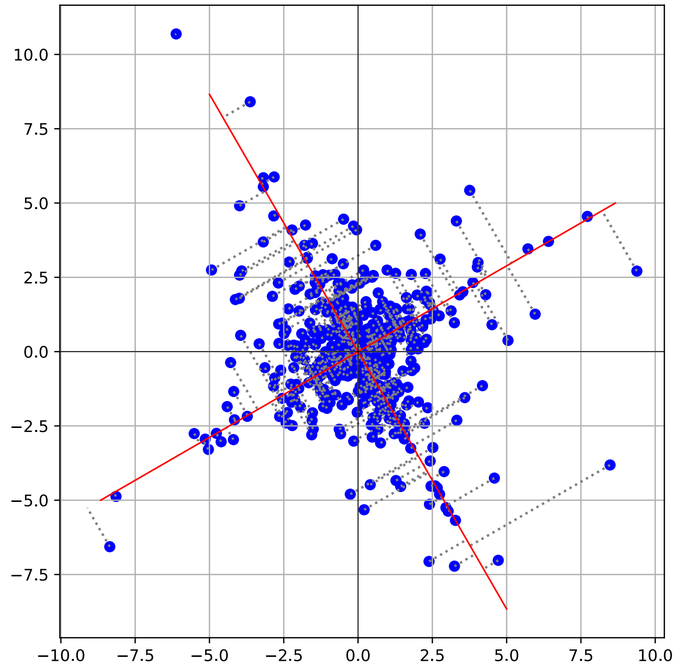
we pursue solving

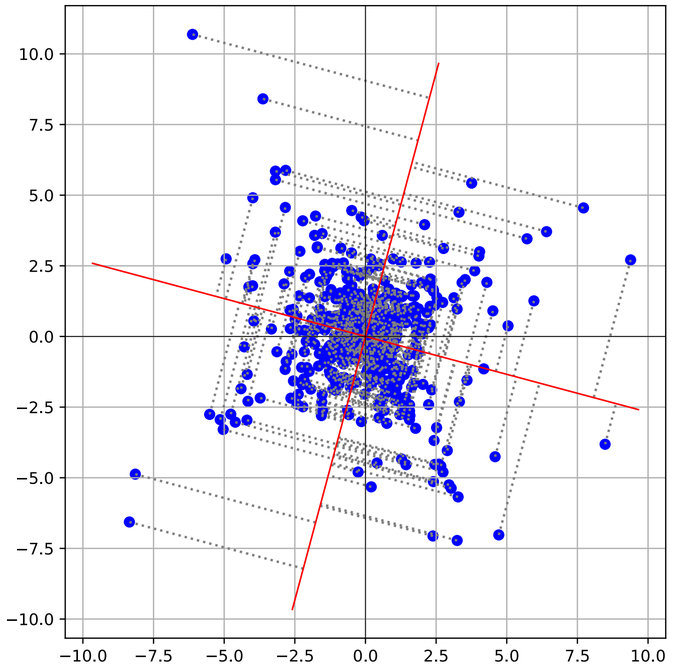
$$\min_{\nu \in \mathscr{P}_\times} W_2^2(\mu, \nu)$$

In principle, this is an **infinite-dimensional** problem, BUT…









Reduction to finite dimension

Let's consider instead

$\mathrm{O}(d) := \{U \in \mathbb{R}^{d\times d} : U^T U = I_{d\times d}\}$ and equivalently formulate

$$\min_{U\in \mathrm{O}(d)} W_2^2(\mu, \mathrm{proj}_{\times U}\#\mu)$$

As we've seen, in practice, we are given $x_1, \ldots, x_n \sim \mu$ i.i.d., so… we aim to solve

$$\min_{U\in \mathrm{O}(d)} \sum_{i=1}^{n} \|X_i - \mathrm{proj}_{\times U}(X_i)\|^2$$

Optimize on a manifold?

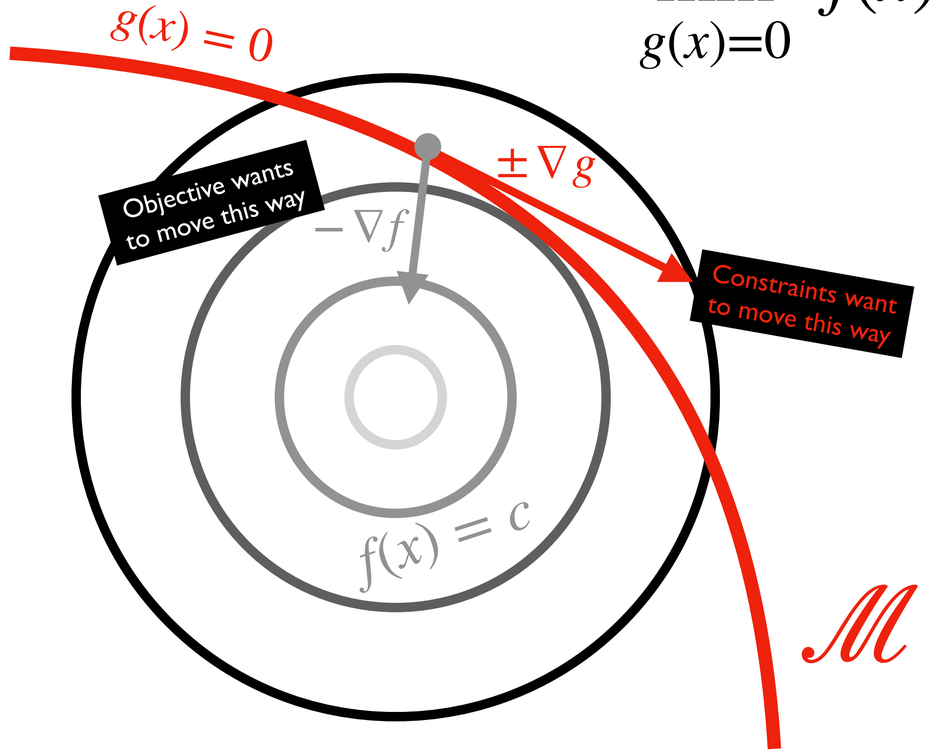
Interlude: Optimization on Manifolds

$$\min_{g(x)=0} f(x)$$

Manifold: $\mathrm{O}(d) := \{U \in \mathbb{R}^{d\times d} : U^T U = I_{d\times d}\}$

Tangent Space: $T_U\mathrm{O}(d) := \{U\Omega : \Omega \in \mathrm{Skew}(d)\} = U\mathrm{Skew}(d)$

Retraction: $R_U(X) := (U+X)(I+X^TX)^{-1/2}$



How to optimize without stepping off of the manifold?

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

"Euclidean gradient descent"

walk direction

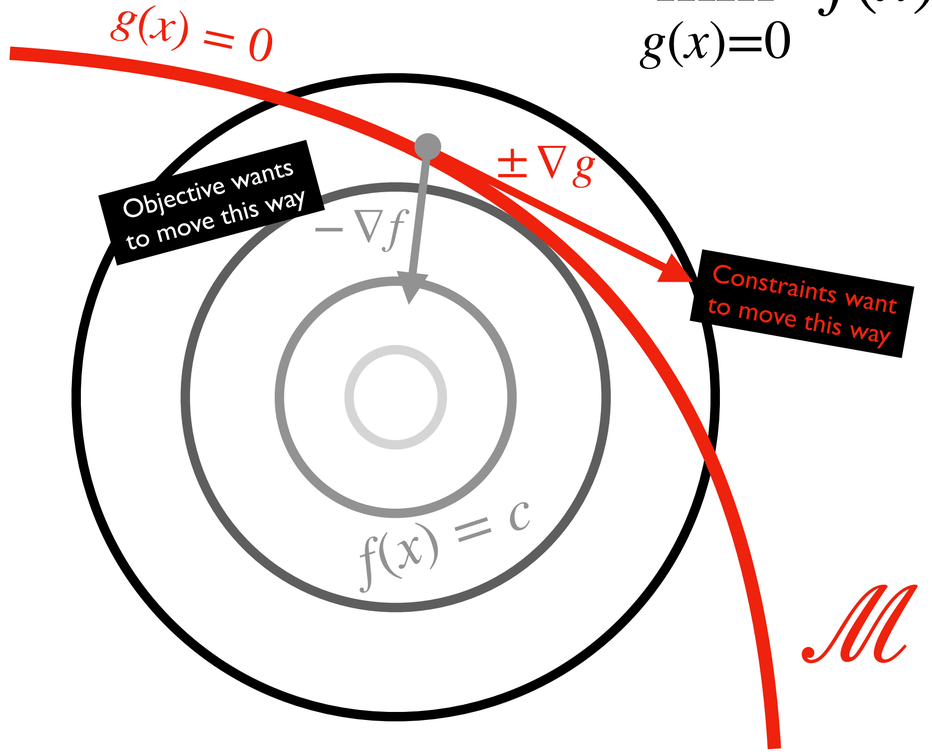
"Manifold gradient descent"

$$x_{k+1} = \exp_{x_k}(-\alpha_k \nabla f(x_k))$$

replaced by a weaker notion: **retraction**

Interlude: Optimization on Manifolds

$$\min_{g(x)=0} f(x)$$



Manifold: $\mathrm{O}(d) := \{U \in \mathbb{R}^{d\times d} : U^T U = I_{d\times d}\}$

Tangent Space: $T_U\mathrm{O}(d) := \{U\Omega : \Omega \in \mathrm{Skew}(d)\} = U\mathrm{Skew}(d)$

Retraction: $R_U(X) := (U + X)(I + X^T X)^{-1/2}$

Riemannian Gradient descent!

$$\min_{U\in\mathrm{O}(d)} W_2^2(\mu, P_U\#\mu)$$

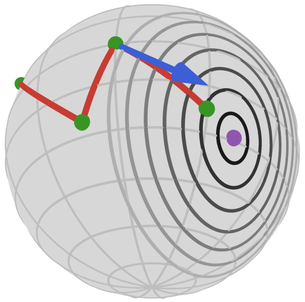
$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$ "Euclidean gradient descent"

walk direction

"Manifold gradient descent"

$$x_{k+1} = \exp_{x_k}(-\alpha_k \nabla f(x_k))$$

replaced by a weaker notion: **retraction**



pymanopt

Convergence of the empirical minimizer

Theorem I:

Suppose μ is absolutely continuous with finite fourth moment, has sub-exponential squared norms, and satisfies an anti-concentration condition away from the optimal frame axes. Suppose further that frames with sufficiently small population risk $\mathscr{E}(U) := W_2^2(\mu, \mathrm{proj}_{\times U}\#\mu)$ lie within a geodesically convex neighborhood $\mathscr{N}$ of the population minimizer U^*. Then any empirical minimizer $U_n^* \in \mathrm{argmin}_{U\in \mathrm{O}(d)} \mathscr{E}_n(U)$ satisfies, apart from a distribution-dependent constant, with high probability,

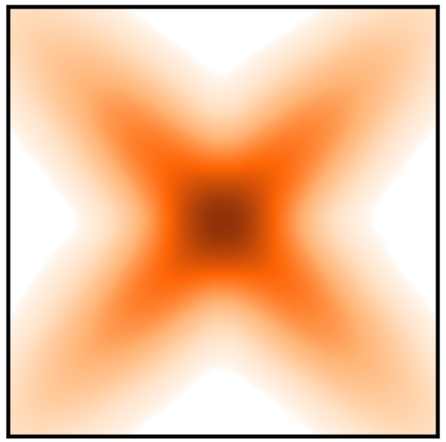
$$d_g(U_n^*, U^*) = \mathscr{O}_p\left(\frac{d^{5/2}\log n}{\sqrt{n}}\right).$$

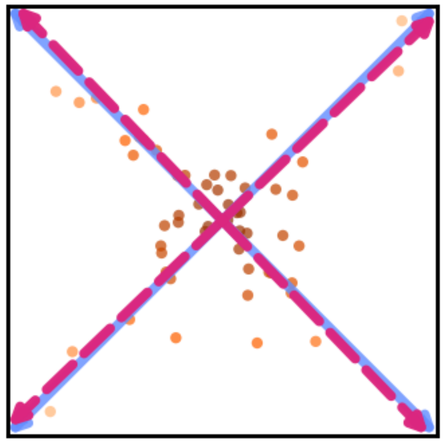
To achieve accuracy ε we need $n = \Omega\left(d^5\varepsilon^{-2}\log^2 n\right)$.

Standard M-estimation steps:

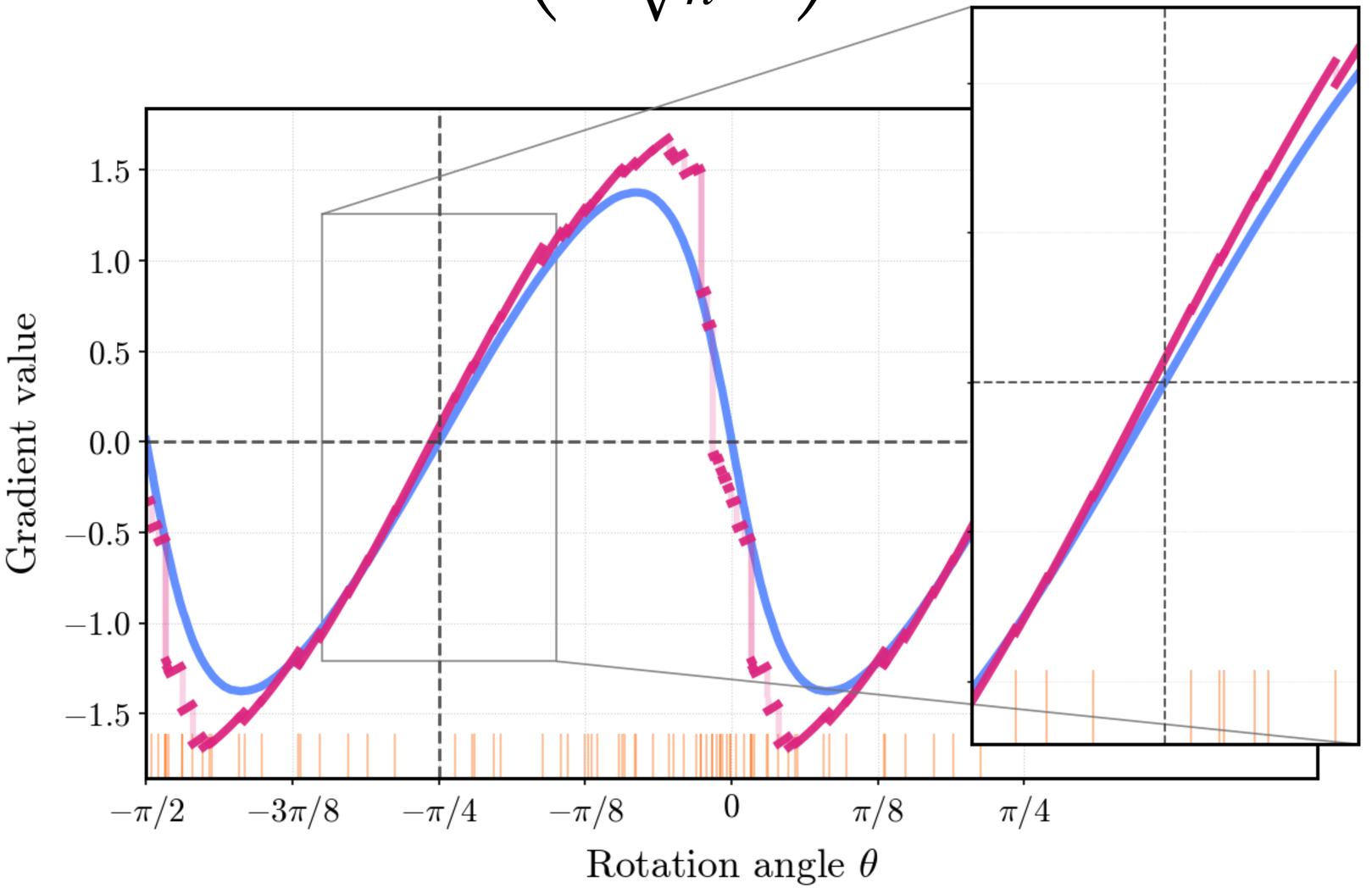
- CLT gives you $O_p(n^{-1/2})$ control of the gradient error at the true minimizer;
- Lipschitz continuity of the gradient lets you extend that pointwise bound to a whole neighborhood, so it also applies at the random, data-dependent empirical minimizer;
- strong convexity converts small gradient into small parameter error.

$$d_g(U_n^*, U^*) = \mathcal{O}_p\left(\frac{d^{5/2}\log n}{\sqrt{n}}\right).$$





Population minimizer
Empirical minimizer

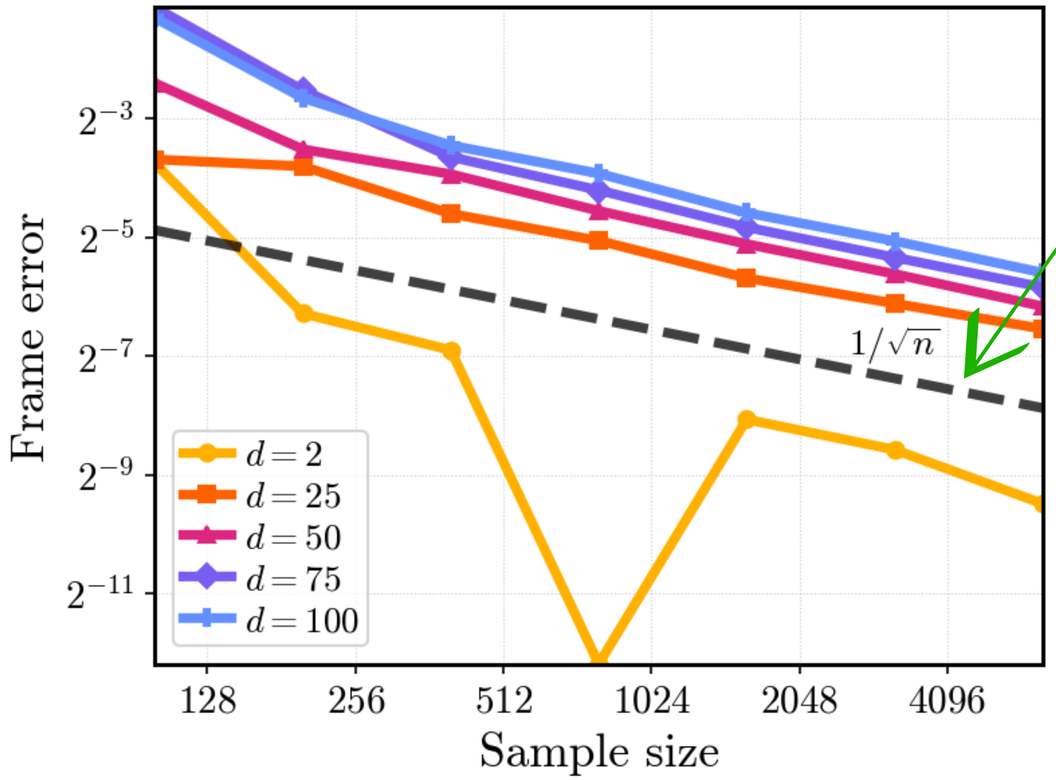


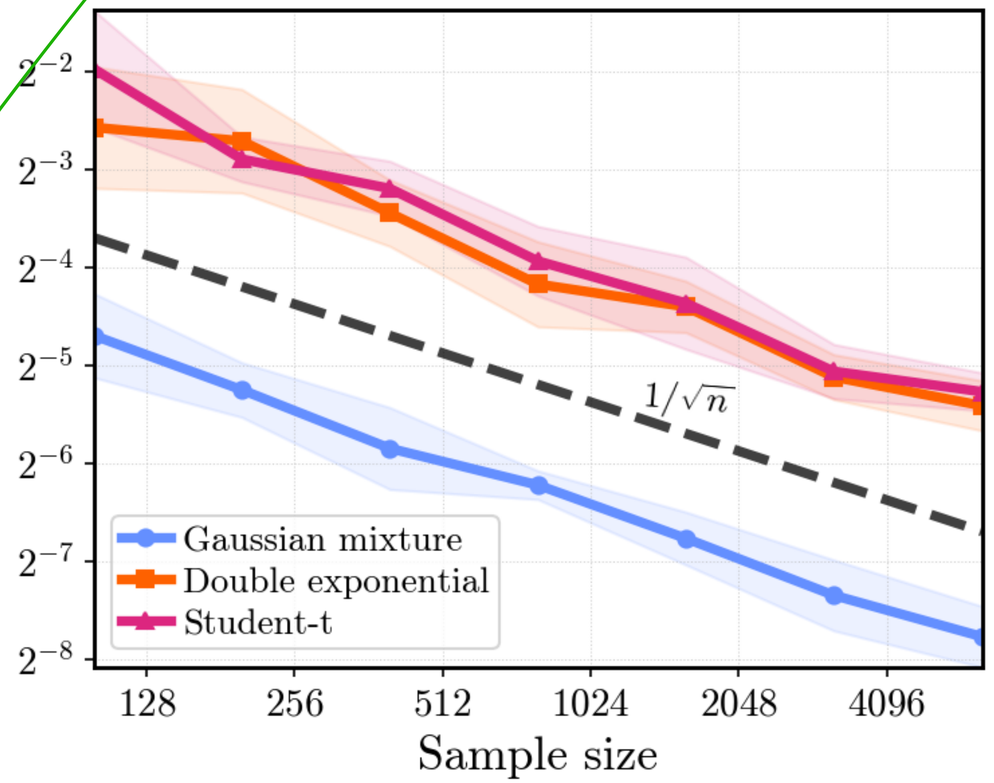
$\nabla^R \mathscr{E}(U)$ $\partial^R \mathscr{E}_n(U)$

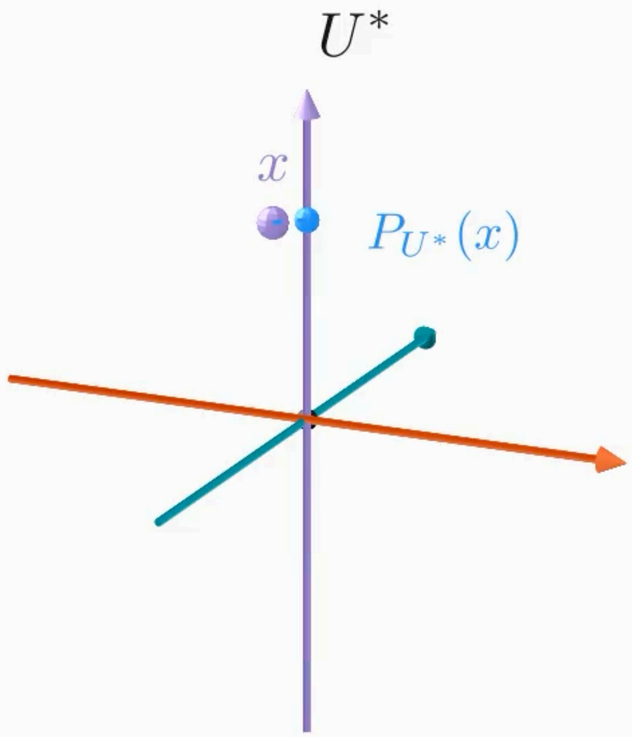
How?

$$d_g(U_n^*, U^*) = \mathcal{O}_p\left(\frac{d^{5/2}\log n}{\sqrt{n}}\right).$$

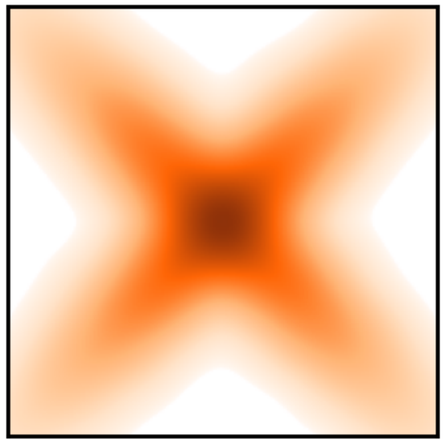
No curse of dimensionality!

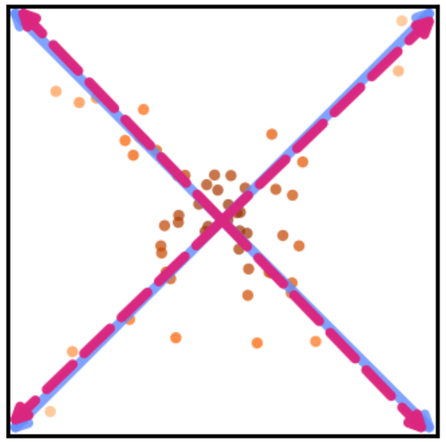




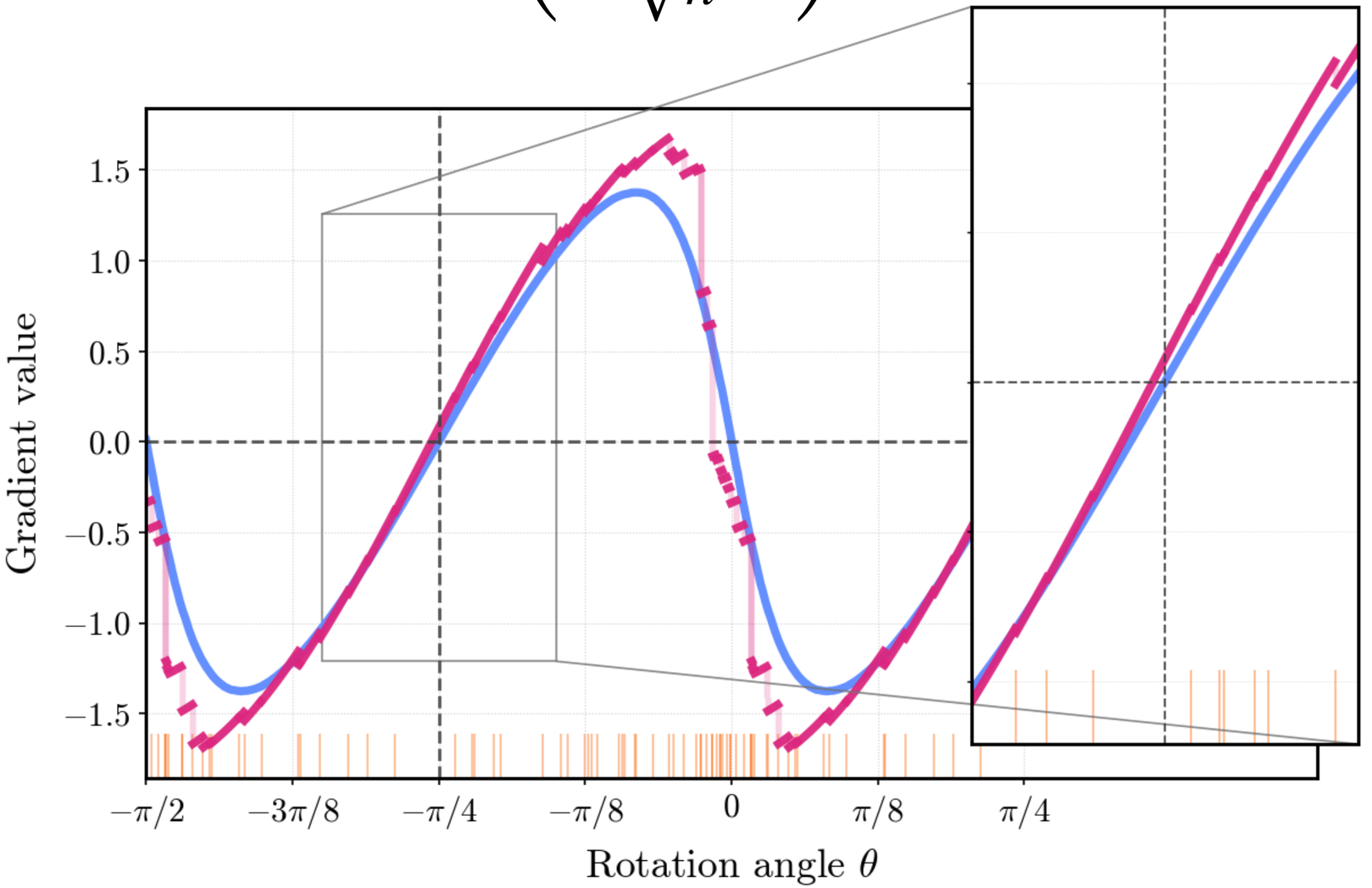

U^*
x
$P_{U^*}(x)$

$$d_g(U_n^*, U^*) = \mathcal{O}_p\left(\frac{d^{5/2}\log n}{\sqrt{n}}\right).$$





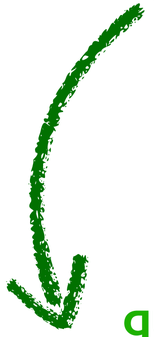
Population minimizer
Empirical minimizer



$\nabla^R \mathscr{E}(U)$ $\partial^R \mathscr{E}_n(U)$

Standard M-estimation steps:

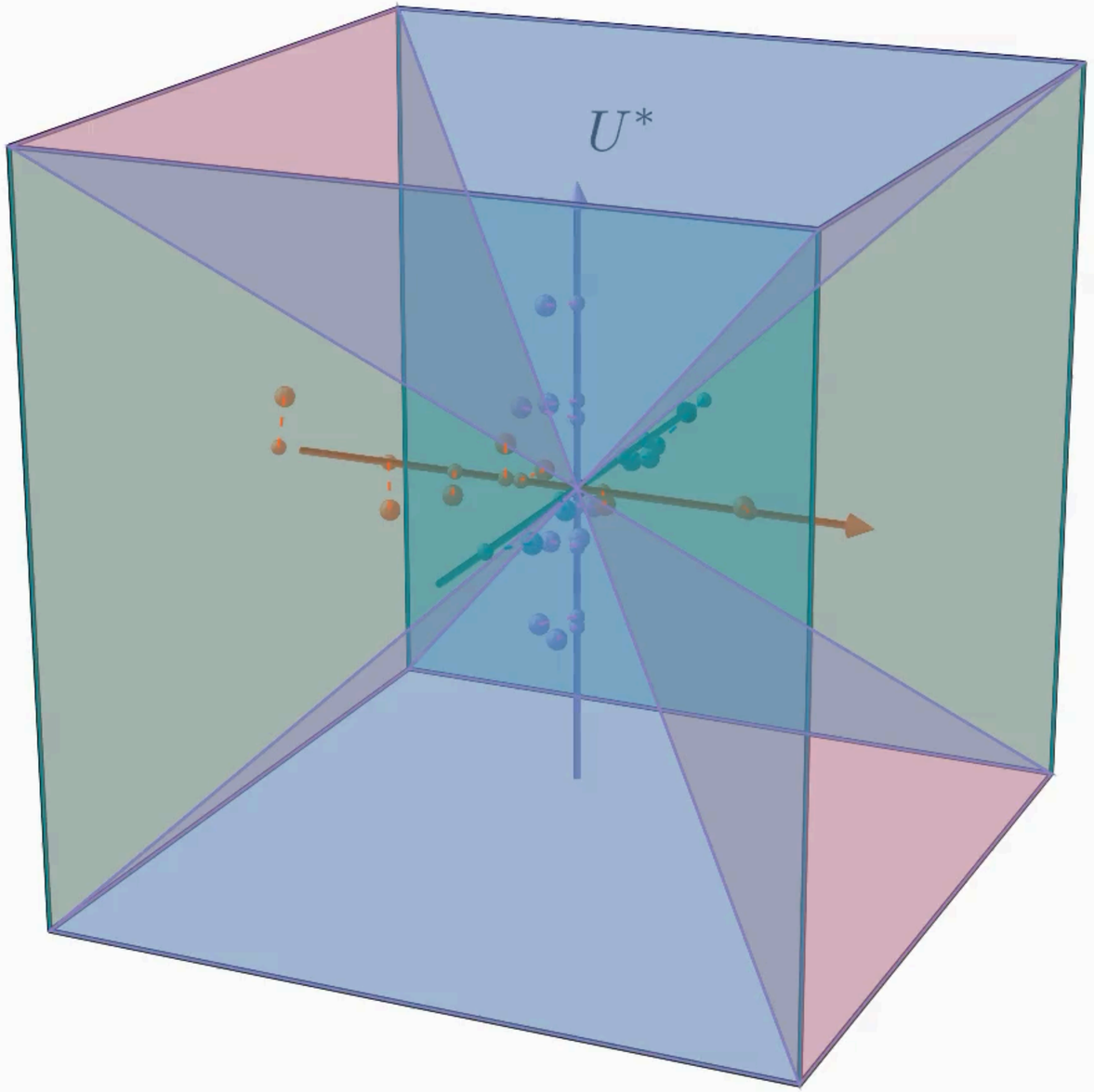
- **CLT gives you $O_p(n^{-1/2})$ control of the gradient error at the true minimizer;**
- **Lipschitz continuity of the gradient lets you extend that pointwise bound to a whole neighborhood, so it also applies at the random, data-dependent empirical minimizer;**
- **strong convexity converts small gradient into small parameter error.**

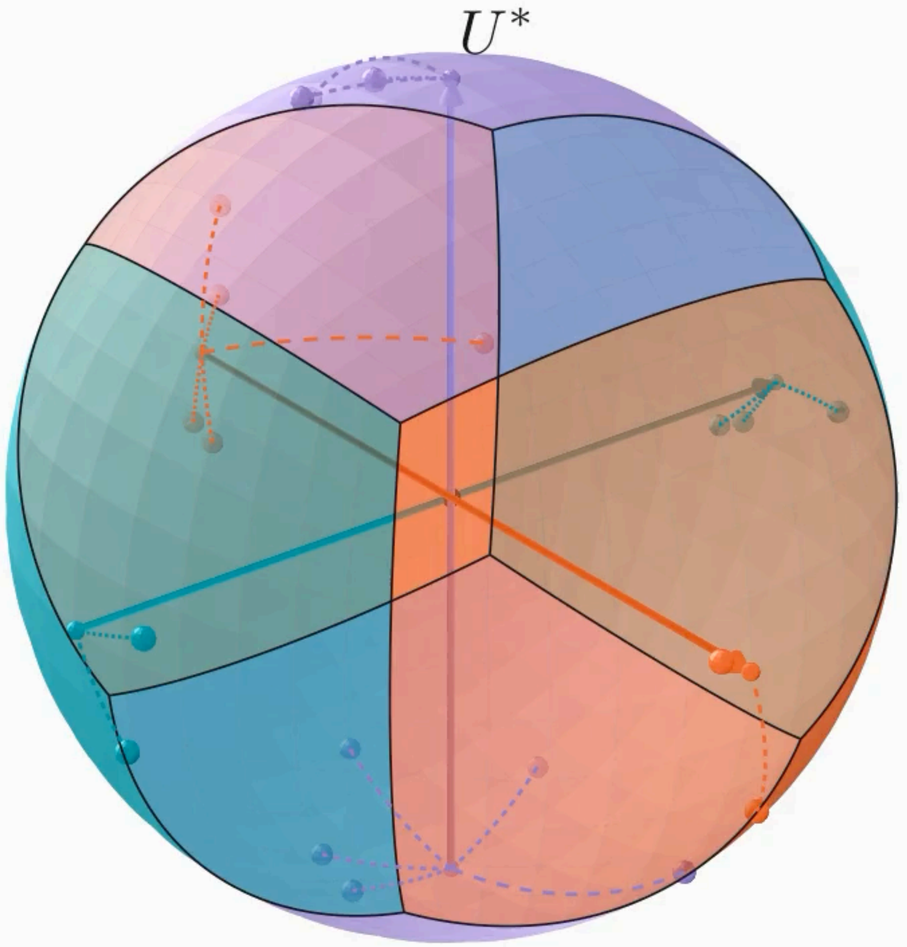


quasi-Lipschitzness

$$\|G_n(U_1) - G_n(U_2)\|_{\mathrm{Fr}} \le \underbrace{4\|x\|_2^2 \cdot \|U_1 - U_2\|_{\mathrm{Fr}}}_{\text{ordinary Lipschitz term}} + \underbrace{Cd(d-1)\frac{\omega_{d-1}}{\omega_d}\frac{\log n}{\sqrt{n}}}_{\text{resolution error (switching)}} \quad \textbf{w/} \quad \|U_1 - U_2\|_{\mathrm{Fr}} \le 1/\sqrt{n}$$

$$\underbrace{C\,d(d-1)\frac{\omega_{d-1}}{\omega_d}\frac{\log n}{\sqrt{n}}}_{\text{resolution error}} = \underbrace{\frac{1}{n}}_{\text{average over } n} \times \underbrace{N_{1/\sqrt{n}}(U_1, U_2)}_{\#\text{ samples that switch axis}} \times \underbrace{O(\log n)}_{\substack{\max \text{ jump per switch,} \\ \|x_i\|^2 \lesssim \log n}}$$


U^*


U^*

Theorem [Switching count concentration]:

For any $U_1, U_2 \in \mathcal{N}$ such that $\|U_1 - U_2\|_{\mathrm{Fr}} \le 1/\sqrt{n}$, with high probability,

$$N_{1/\sqrt{n}}(U_1, U_2) = d(d-1)\left(C\frac{\omega_{d-1}}{\omega_d}\sqrt{n} + \cdots \right)$$

where ω_d denotes the surface area of the unit sphere in $\mathbb{R}^d$.

$$\underbrace{C\, d(d-1)\frac{\omega_{d-1}}{\omega_d}\frac{\log n}{\sqrt{n}}}_{\text{resolution error}} = \underbrace{\frac{1}{n}}_{\text{average over } n} \times \underbrace{N_{1/\sqrt{n}}(U_1, U_2)}_{\#\text{ samples that switch axis}} \times \underbrace{O(\log n)}_{\substack{\text{max jump per switch,} \\ \|x_i\|^2 \lesssim \log n}}$$

$$d_g(U_n^*, U^*) \leq \frac{1}{m}\left\|\nabla_R \mathscr{E}(U_n^*)\right\|_F \qquad \text{(geodesic strong convexity)}$$

(uniformity argument)
$$\leq \frac{1}{m}\sup_{U\in\mathcal{N}}\left\|\nabla_R \mathscr{E}(U) - G_n(U)\right\|_F$$

(ε-net argument)
$$\leq \frac{1}{m}\left(\max_{U_k}\|\nabla_R \mathscr{E}(U_k) - G_n(U_k)\|_F + \max_{U_k}\sup_{\|U-U_k\|\leq\delta}\|G_n(U) - G_n(U_k)\|_F\right)$$

(pointwise CLT + quasi-L)
$$\leq \frac{1}{m}\left(C_1 d\sqrt{\frac{Cd(d-1)\log n}{n}} + 4\|x\|_{L^2}^2\delta + C_2 d(d-1)\frac{\omega_{d-1}}{\omega_d}\frac{\log n}{\sqrt{n}}\right)$$

(absorbing constants)
$$\leq \frac{1}{m}\left(C_1 d^2\sqrt{\frac{\log n}{n}} + C_2 d(d-1)\frac{\omega_{d-1}}{\omega_d}\frac{\log n}{\sqrt{n}}\right)$$

$$= O_p\left(\frac{d^{5/2}\log n}{\sqrt{n}}\right)$$

Computational guarantees

Theorem II:

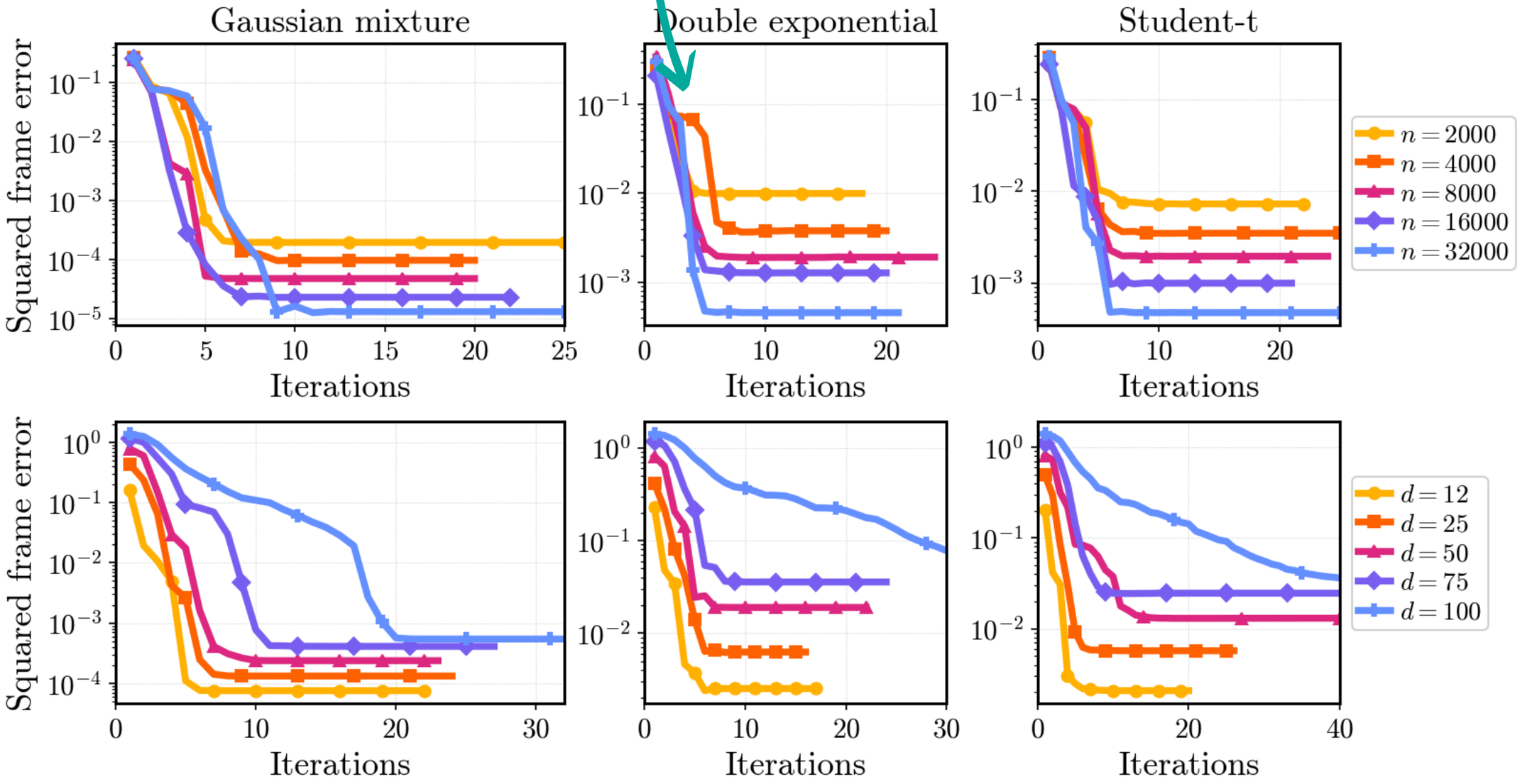
Suppose μ is absolutely continuous with finite fourth moment and that the population risk $\mathscr{E}(U)$ is geodesically strongly convex on $\mathscr{N}$. The Riemannian online-SGD algorithm on $\mathrm{O}(d)$ with step size $\eta \in (0, 2m/L^2)$, run for $T = \mathrm{O}(\log n)$ iterations, initialized at $U_0 \in \mathscr{N}$, produces a frame U_T satisfying

$$\mathbb{E}\left[d_g(U_T, U^*)^2\right] = \mathrm{O}\left(\frac{\log(n)^3}{n}\right).$$

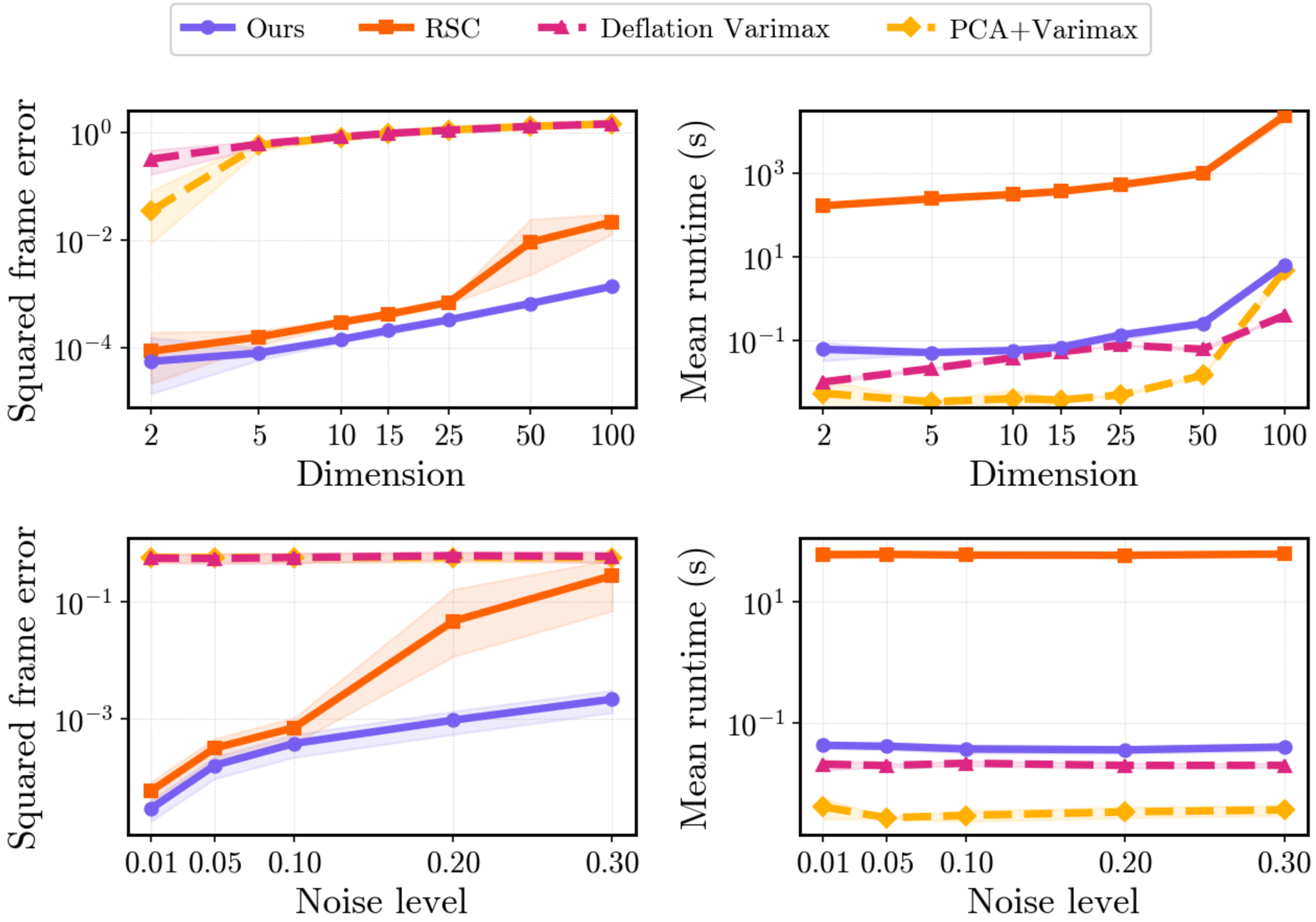
Under the additional condition that the squared norms are sub-exponential, the same rate holds with high probability, with an exponentially small failure probability. The constants m and L quantify the strong convexity and smoothness of the population objective $\mathscr{E}$ near U^*.

$$\mathbb{E}\left[d_g(U_T, U^*)^2\right] = \mathrm{O}\left(\frac{\log(n)^3}{n}\right).$$

Steep initial slope

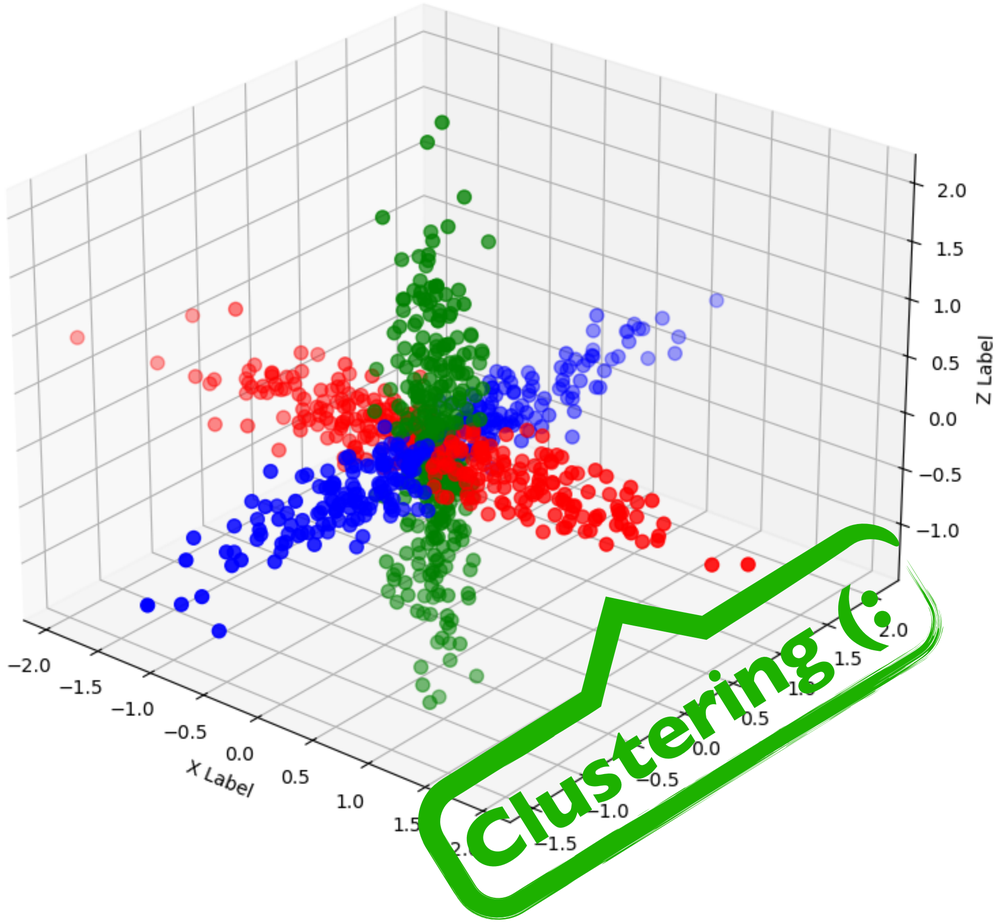


Compared to other methods...

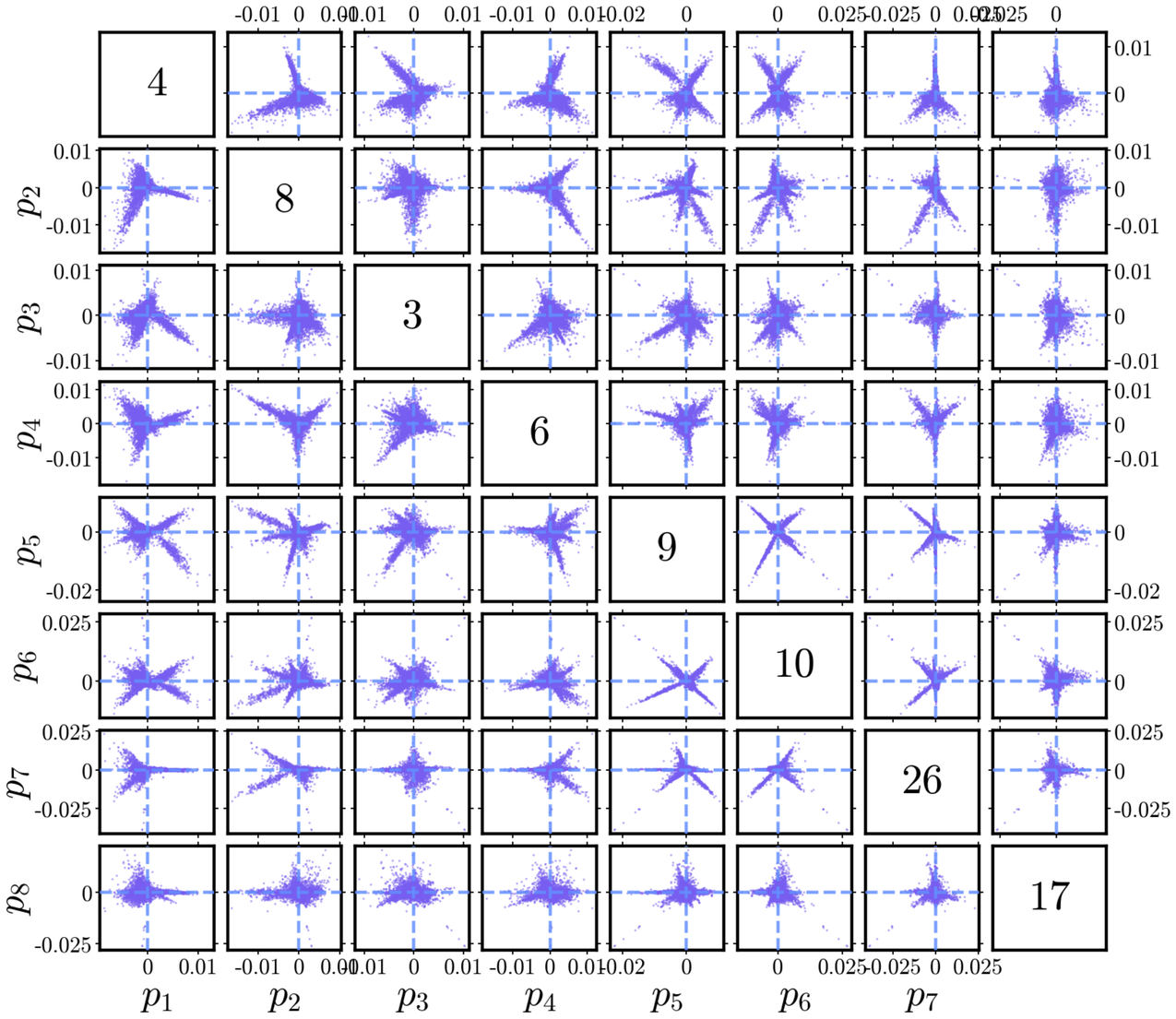


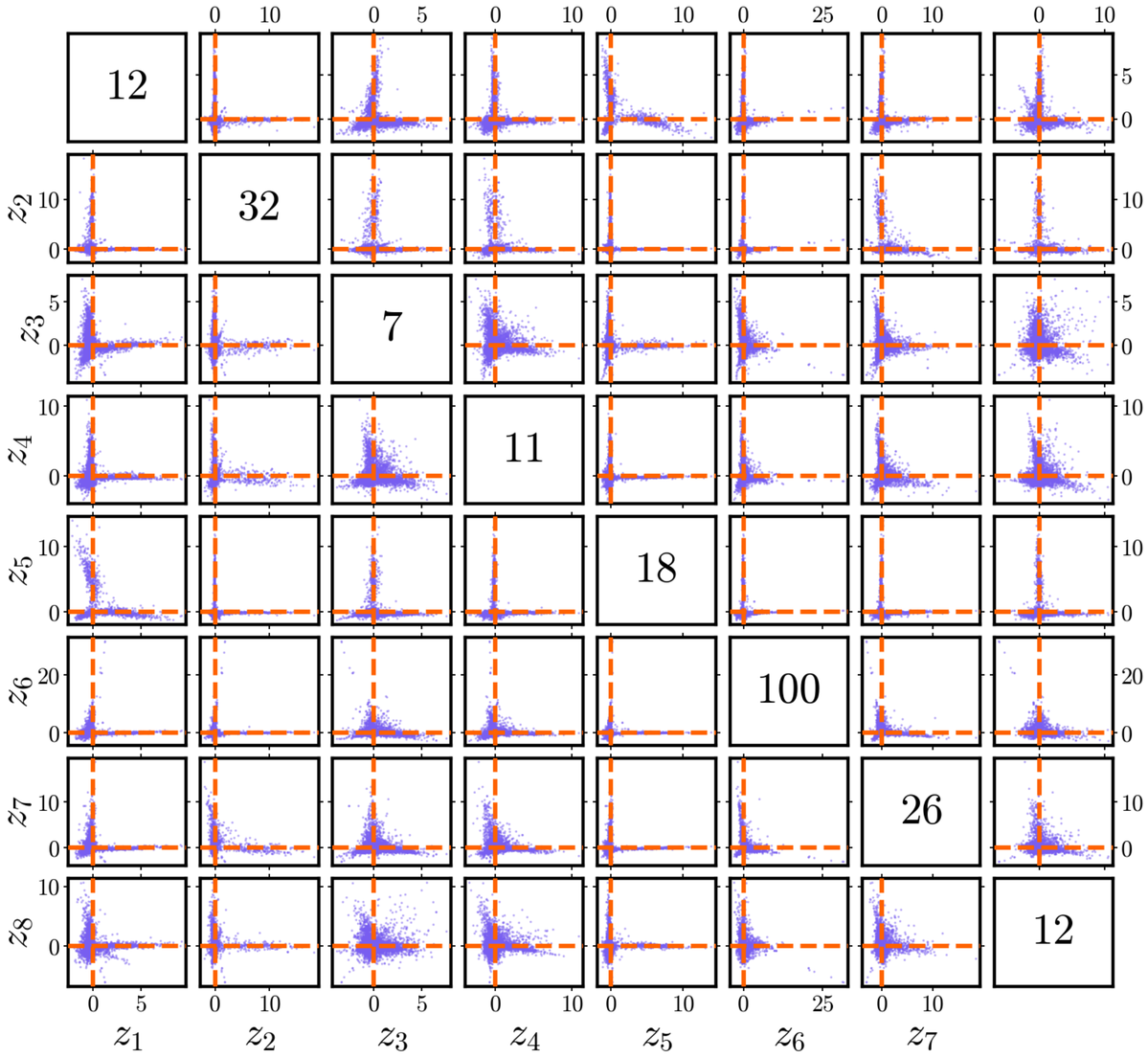
Applications

Given $X \in \mathbb{R}^{d \times n}$, ***recover the labels*** $y \in \{1, \ldots, k\}$



NYT bag of words dataset

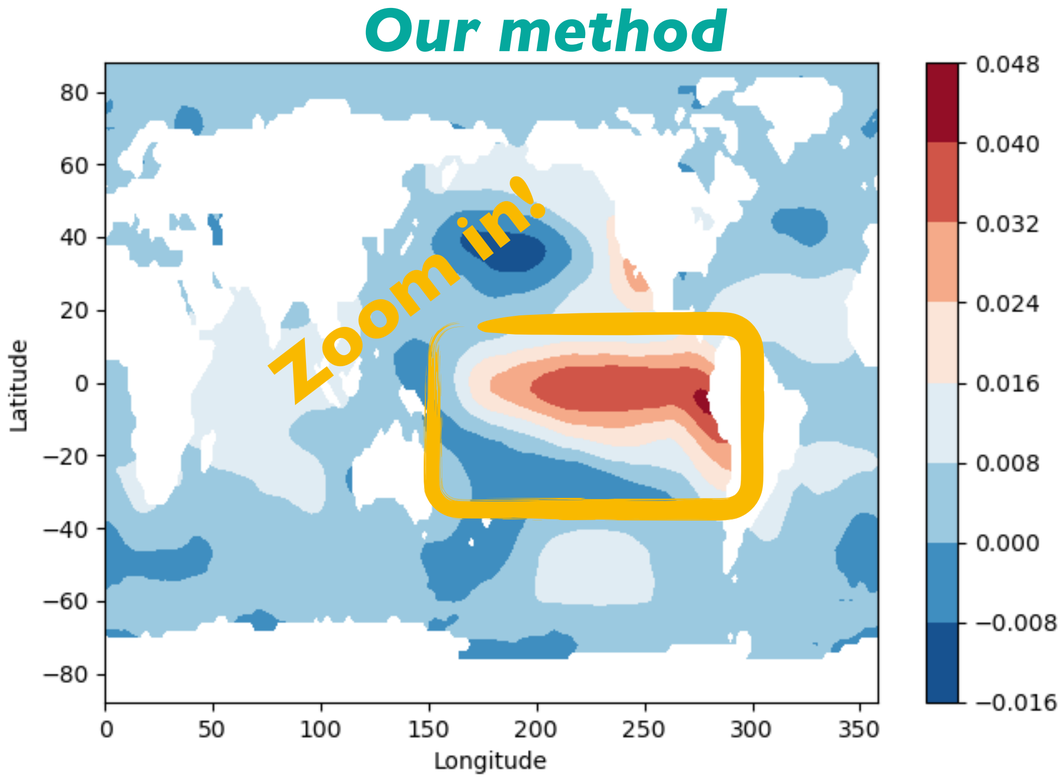




EOF Analysis

Climate researchers want to study **spatial variables** and how they change with time (multidimensional time series data)

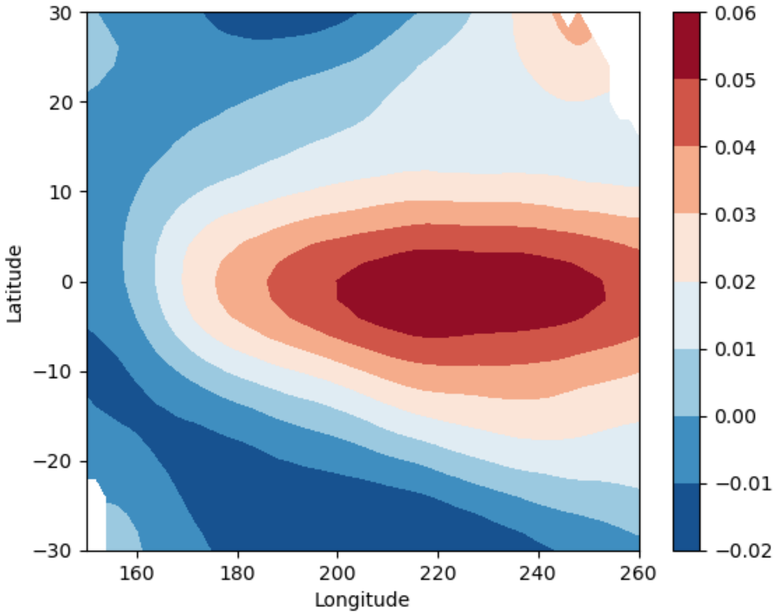
Goal: Fewer **uncorrelated** latent variables ***(factors)*** of an **anomaly covariance matrix**

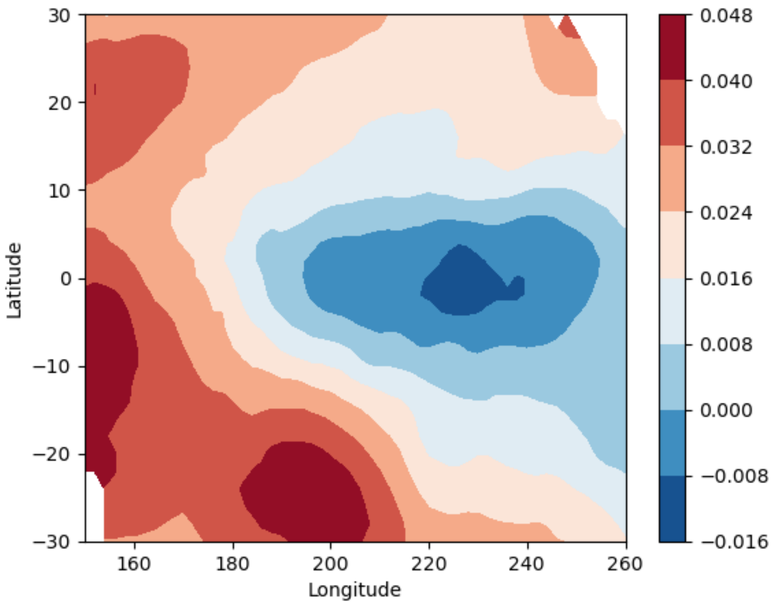


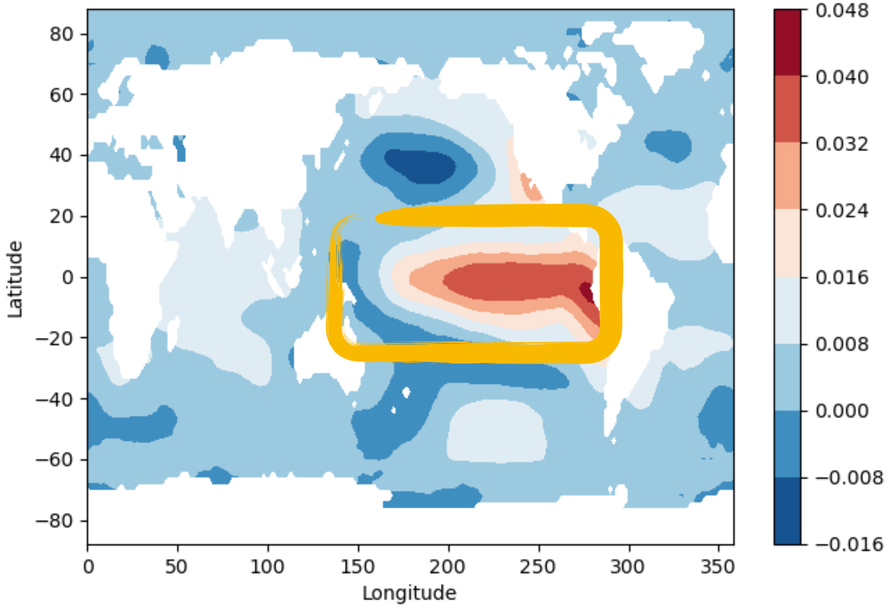
Observed field:

$$X = [X_1, \ldots, X_d]^\top, \text{ with } X_t = \left[X_{t_1}, \ldots, X_{t_d}\right] \implies \mathrm{Cov}(X) = \mathbb{E}XX^\top$$

(Real) Real-data applications



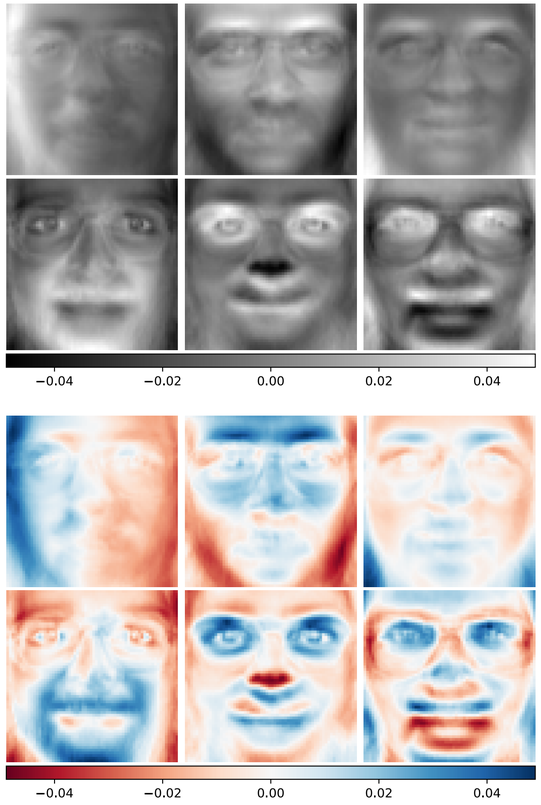




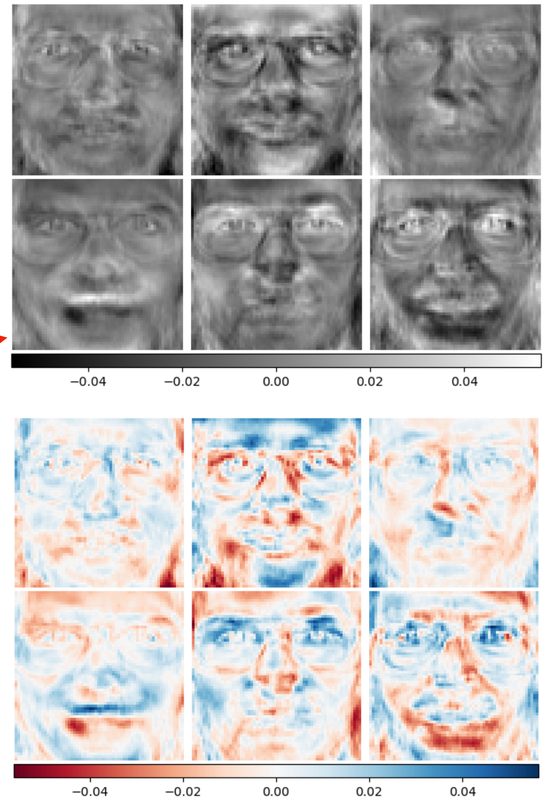
In **sea surface temperature** data, GFA detects **El Niño** warming in the central and eastern Pacific, while orthogonality captures **La Niña**-like cool water patterns.

Faces dataset

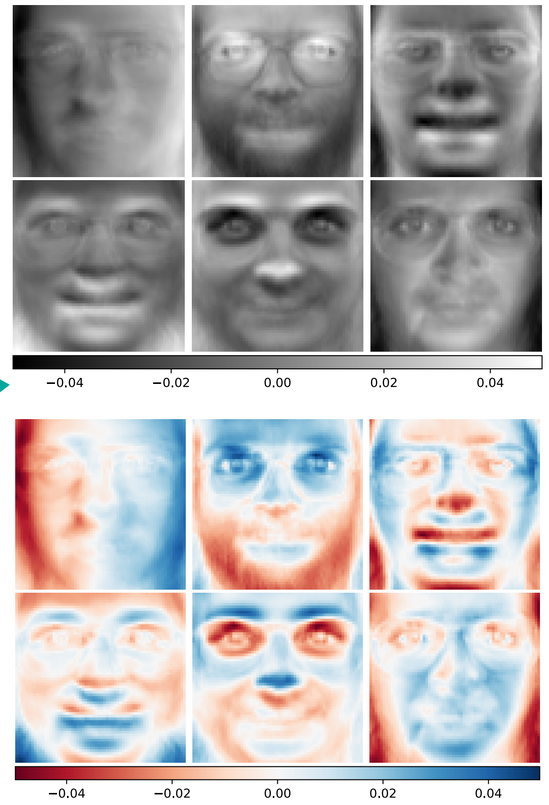
PCA



Varimax



GFA



GFA provides a **more interpretable** factor rotation of high-dimensional datasets **compared to Varimax**

Summary

- Constructed a method that uncovers latent (fewer) factors that are a linear combination of correlated variables
- This method is scalable and achieves better statistical performance than other classical methods
- Provided optimization guarantees for this method
- Demonstrate its applicability in real-world datasets

Gracias!!